DPhil Thesis

Partial Identifiability and Misspecification in Inverse Reinforcement Learning

Joar Skalse

Saint Edmund Hall

2025

Submitted in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

Supervised by Professor Alessandro Abate

Acknowledgements

First and foremost, I would like to thank my supervisor, Alessandro Abate, for his mentorship, guidance, backing, and support (both personal and academic). I especially want to thank him for his support while I fell ill in 2021 — his help during that time was incredibly important, and will not be forgotten. I also want to thank him for believing in me, and for taking me on as a student — I hope I have proven that he made the right decision. Finally, I also want to thank him for being an incredibly nice, thoughtful, and friendly person — studying under him has been a great experience! The work in this thesis would not have been possible without him, and for that I owe him my sincerest gratitude.

In addition to Alessandro, I also want to thank everyone I collaborated with during my DPhil. For my core research, this includes (in no particular order) Matthew Farrugia-Roberts, Adam Gleave, Stuart Russell, Lucy Farnik, Sumeet Ramesh Motwani, Erik Jenner, Nikolaus Howe, Dmitrii Krasheninnikov, and David Krueger. For my non-core research undertaken in parallel with my DPhil research, this also includes (in no particular order) Charlie Griffin, Jacek Karwowski, Oliver Hayman, Xingjian Bai, Klaus Kiendlhofer, Rohan Subramani, Marcus Williams, Max Heitmann, Halfdan Holm, Lukas Fluri, Leon Lang, Patrick Forré, David 'davidad' Dalrymple, Yoshua Bengio, Max Tegmark, Sanjit Seshia, Steve Omohundro, Christian Szegedy, Ben Goldhaber, Nora Ammann, Joe Halpern, Clark Barrett, Ding Zhao, Tan Zhi-Xuan, Jeannette Wing, and Joshua Tenenbaum. All of my research collaborations were very good experiences, from which I learnt a lot. I can only hope that I was a good collaborator as well.

In addition to my collaborators, I also want to thank the many people with whom I have had interesting and stimulating conversations about my research (and research more generally) throughout the course of my DPhil. In addition to many of those already mentioned above, this includes (in no particular order) Chris van Merwijk, Caspar Oesterheld, Lewis Hammond, Joe Benton, Ondrej Bajgar, Hjalmar Wijk, Ryan Carey, Michael Cohen, Richard Ngo, Vojtech Kovarik, Daniel Filan, Lauro Langosco, Stuart Armstrong, Nandi Schoots, Sam Clarke, Jan Brauner, Rohin Shah, Daniel Murfet, Alexander Gietelink Oldenziel, Liam Carroll, Jesse Hoogland, Justin Shovelain, Hugo Berg, Vladimir Mikulik, Vivek Hebbar, Ian McKenzie, Peter Barnett, Tamera Lanham, Thomas Kwa, James Fox, Alyosha Latyntsev, and many others, who all helped me to refine my thinking, see problems from new perspectives, discover new ideas, find solutions to problems, or spot mistakes in my thinking.

I highly value the conversations we have had, and I look forward to many more conversations in the future, as the field of artificial intelligence continues to progress.

I also want to thank all the (former) members of the Future of Humanity Institute, including those not already mentioned above. The FHI was a very special place, the likes of which we may never see again. The people at the FHI all contributed to creating a very stimulating intellectual environment, which certainly helped me become a better researcher. I also want to thank the FHI, as an institution, for providing the funding that supported this thesis, and thereby making this research possible in the first place. The FHI may not have been able to ensure its own survival, but I believe it did help with ensuring the survival of humanity.

Finally, I also want to thank my parents, Mats Koel and Inger Skalse, my brother, Emil Skalse, and my fiancée, Anna Keszthelyi. They have all supported me in more ways than I could possibly enumerate here, and I could not have done this without them.

To all of you,

Joar Skalse

Statement of Originality

All work presented in this thesis is my individual work, unless otherwise stated. Professor Alessandro Abate provided comments and feedback on the writing in this thesis (and provided supervision during the research that this thesis encompasses). My fiancée, Anna Keszthelyi, also helped me with the creation of some of the explanatory figures that appear in this text.

Most of the material in this dissertation has appeared in published work, specifically in *Defining and Characterising Reward Hacking* (Skalse, Howe, et al., 2022), *Invariance in Policy Optimisation and Partial Identifiability in Reward Learning* (Skalse, Farrugia-Roberts, et al., 2022), *Misspecification in Inverse Reinforcement Learning* (Skalse and Abate, 2023a), *STARC: A Gemeral Framework For Quantifying Differences Between Reward Functions* (Skalse, Farnik, et al., 2023), and *Quantifying the Sensitivity of Inverse Reinforcement Learning to Misspecification* (Skalse and Abate, 2024). The correspondence between the material in this thesis and the material in these earlier publications is described in Section 1.3. However, in each case, the material presented in this thesis is still entirely my own work, with the other authors of these aforementioned research papers being responsible for other material in those papers (which does not appear in this thesis). The only exception to this is some proofs from *Invariance in Policy Optimisation and Partial Identifiability in Reward Learning*, which were derived by me together with Matthew Farrugia-Roberts, in such a way that neither of us can be given sole credit for these theorems. This concerns Propositions 29-32, Lemmas 67-71, Proposition 72, and Theorems 73-75, found in Sections 4.1, 5.1, and 5.2. All other theorems appearing in this thesis were proven entirely by me.

Abstract

The aim of Inverse Reinforcement Learning (IRL) is to infer a reward function R from a policy π. This problem is difficult, for several reasons. First of all, there are typically multiple reward functions which are compatible with a given policy; this means that the reward function is only *partially identifiable*, and that IRL contains a certain fundamental degree of ambiguity. Secondly, in order to infer R from π, an IRL algorithm must have a *behavioural model* that describes how π relates to R. However, the true relationship between human preferences and human behaviour is very complex, and practically impossible to fully capture with a simple model. This means that the behavioural model in practice will be *misspecified*, which raises the worry that it might lead to unsound inferences if applied to real-world data. In this thesis, we provide a comprehensive mathematical analysis of partial identifiability and misspecification in IRL. Specifically, we fully characterise and quantify the ambiguity of the reward function under all of the behavioural models that are most common in the current IRL literature. We also provide necessary and sufficient conditions that describe precisely how the observed demonstrator policy may differ from each of the standard behavioural models before that model leads to faulty inferences about the reward function R. In addition to this, we introduce a cohesive framework for reasoning about partial identifiability and misspecification in IRL, together with several formal tools that can be used to easily derive the partial identifiability and misspecification robustness of new IRL models, or analyse other kinds of reward learning algorithms.

List of Publications

Main Publications:

1. Defining and Characterizing Reward Hacking. **Joar Skalse**, Nikolaus Howe, Dmitrii Krasheninnikov, David Krueger. Published at NeurIPS 2022.

2. Invariance in Policy Optimisation and Partial Identifiability in Reward Learning. **Joar Skalse***, Matthew Farrugia-Roberts*, Stuart Russell, Alessandro Abate, Adam Gleave. *Equal contribution. Published at ICML 2023.

3. Misspecification in Inverse Reinforcement Learning. **Joar Skalse**, Alessandro Abate. Published at AAAI 2023. This paper was awarded the recognition of **Outstanding Paper** at AAAI for "exemplify[ing] the highest standards in technical contribution and exposition".

4. STARC: A General Framework For Quantifying Differences Between Reward Functions. **Joar Skalse**, Lucy Farnik, Sumeet Ramesh Motwani, Erik Jenner, Adam Gleave, Alessandro Abate. Published at ICLR 2024.

5. Quantifying the Sensitivity of Inverse Reinforcement Learning to Misspecification. **Joar Skalse**, Alessandro Abate. Published at ICLR 2024.

Other Relevant Publications:

6. Reinforcement Learning in Newcomblike Environments. James Bell*, Linda Linsefors*, Caspar Oesterheld*, **Joar Skalse***. *Equal contribution. Published at NeurIPS 2021.

7. Is SGD a Bayesian sampler? Well, almost. Chris Mingard, Guillermo Valle-Pérez, **Joar Skalse**, Ard Luis. Published in JMLR 2021.

8. Lexicographic Multi-Objective Reinforcement Learning. **Joar Skalse**, Lewis Hammond, Charlie Griffin, Alessandro Abate. Published at IJCAI 2022.

9. All's Well That Ends Well: Avoiding Side Effects with Distance-Impact Penalties. Charlie Griffin, **Joar Skalse**, Lewis Hammond, Alessandro Abate. Published at the NeurIPS 2022 Safety Workshop.

10. On the Limitations of Markovian Rewards to Express Multi-Objective, Risk-Sensitive, and Modal Tasks. **Joar Skalse**, Alessandro Abate. Published at UAI 2023.

11. Goodhart's Law in Reinforcement Learning. Jacek Karwowski, Oliver Hayman, Xingjian Bai, Klaus Kiendlhofer, Charlie Griffin, **Joar Skalse**. Published at ICLR 2024.

12. On The Expressivity of Objective-Specification Formalisms in Reinforcement Learning. Rohan Subramani, Marcus Williams, Max Heitmann, Halfdan Holm, Charlie Griffin, **Joar Skalse**. Published at ICLR 2024.

13. Partial Identifiability in Inverse Reinforcement Learning For Agents With Non-Exponential Discounting. **Joar Skalse**, Alessandro Abate. Published at AAAI 2025.

Noteworthy Unpublished Work:

14. Towards Guaranteed Safe AI: A Framework for Ensuring Robust and Reliable AI Systems. David Dalrymple*, **Joar Skalse***, Yoshua Bengio, Stuart Russell, Max Tegmark, Sanjit Seshia, Steve Omohundro, Christian Szegedy, Ben Goldhaber, Nora Ammann, Alessandro Abate, Joe Halpern, Clark Barrett, Ding Zhao, Tan Zhi-Xuan, Jeannette Wing, Joshua Tenenbaum. *Equal contribution. ArXiv preprint, 2024.

15. The Perils of Optimizing Learned Reward Functions: Low Training Error Does Not Guarantee Low Regret. Lukas Fluri, Leon Lang, Alessandro Abate, Patrick Forré, David Krueger, **Joar Skalse**. ArXiv preprint, 2024.

Contents

1 Introduction **1**
1.1 Background and Context 1
1.2 Related Work 4
1.3 Contributions 18

2 Technical Background **20**
2.1 Reinforcement Learning 20
2.2 Inverse Reinforcement Learning 23
2.3 Metrics, Pseudometrics, and Norms 26

3 New Definitions and Formalisms **28**
3.1 Partial Identifiability 28
3.2 Misspecification Robustness 36
3.3 Intermediate Results About Our Definitions 42
3.4 Reward Transformations 51
3.5 Behavioural Models 58

4 Comparing Reward Functions **60**
4.1 Key Properties of Reward Transformations 60
4.2 Equivalent Reward Functions 66
4.3 Unhackable Reward Functions 73
4.4 STARC Metrics 77
4.5 Understanding STARC Metrics 90
4.6 Soundness and Completeness 93

5 Partial Identifiability **112**
5.1 Invariances of Intermediate Objects 112
5.2 Invariances of Policies 115
5.3 Ambiguity Tolerance and Applications 118
5.4 Transfer Learning 122

6 Misspecification With Equivalence Relations **133**
6.1 Necessary and Sufficient Conditions 133
6.2 Wider Classes of Policies 139
6.3 Misspecified Parameters 142
6.4 Transfer Learning 144

7 Misspecification With Metrics **145**
7.1 Necessary and Sufficient Conditions 146
7.2 Perturbation Robustness 152
7.3 Misspecified Parameters 156
7.4 Transfer Learning 159

8 Generalising Our Analysis **160**
8.1 Assumptions About Inductive Bias 160
8.2 Assumptions About the True Reward 166
8.3 Restricting the Space of Reward Functions 168
8.4 Making the Analysis Probabilistic 177
8.5 Stronger Equivalence Conditions 178

9 Discussion **184**
9.1 Impact and Significance 184
9.2 Limitations and Further Work 187

References **189**

All men by nature desire to know.
— Aristotle, sometime in the 4$^{\text{th}}$ century BCE.

1 Introduction

In this chapter, we provide the background and broader context for the work in this thesis, an overview of major related work from the existing literature, an overview of our contributions, and an overview of the structure of this text.

1.1 Background and Context

Inverse Reinforcement Learning (IRL) is an area of machine learning that is concerned with the problem of inferring what *objective* an agent is pursuing based on the *actions* which that agent takes within some environment (Ng and Russell, 2000). IRL can be related to the notion of *revealed preferences* in psychology and economics, since it aims to infer *preferences* from *behaviour* (Rothkopf and Dimitrakakis, 2011). There are many possible applications of IRL. For example, it has been used in natural science contexts, as a tool for understanding animal behaviour (Yamaguchi et al., 2018). It can also be used in various engineering contexts; many important tasks can be represented as sequential decision-making problems, where the goal is to maximise a given *reward function* over several steps (Sutton and A. G. Barto, 2018). However, for many complex tasks it can be very challenging to manually specify a reward function that robustly incentivises the intended behaviour (see e.g. Clark and Amodei, 2016; Paulus, Xiong, and Socher, 2018; Ibarz et al., 2018a;

Manheim and Garrabrant, 2019; Krakovna, Uesato, et al., 2020; Knox et al., 2023; Pang et al., 2022). In those contexts, IRL can be employed to automatically *learn* a good reward function, based on demonstrations of correct behaviour (e.g. Abbeel, Coates, and Ng, 2010; A. Singh et al., 2019). IRL can also be used as a tool for *imitation learning*, where the goal is to use machine learning to clone the behaviour of an agent. In these cases, IRL can improve metrics such as out-of-distribution robustness (e.g. Hussein et al., 2017). Overall, IRL relates to many fundamental questions about goal-directed behaviour and agent-based modelling.

It is important to note that the properties which we desire an IRL method to have will depend on the context in which that IRL method will be applied. For example, when IRL is used as a tool for imitation learning, it is not fundamentally important that the inferred preferences actually correspond to the true intentions of the demonstrator, as long as they help the imitation learning process. However, when IRL is used to understand the preferences and motivations of an agent (as in e.g. Hadfield-Menell et al., 2016, etc), then it is crucial that the inferred preferences actually capture the true intentions of the observed agent as faithfully as possible. We should note that this thesis is written with mainly the latter motivation in mind.

IRL faces several fundamental challenges. First of all, the IRL problem is typically formalised as the problem of inferring a reward function R from a policy π.[1] To do this, an IRL algorithm needs a model of how π relates to R, which is referred to as a *behavioural model*. However, under most behavioural models, there are typically multiple reward functions that are consistent with each given policy. For example, two different reward functions may result in exactly the same optimal policy. In that case, we cannot distinguish between those reward functions by observing their optimal policy. This means that the reward function is ambiguous, or *partially identifiable*, based on this data source. This ought to be intuitive: informally, there can be multiple different reasons for doing something, which means that observed behaviour sometimes can be explained in multiple different ways. For this reason, the IRL problem is fundamentally ambiguous, and we should *prima*

[1]Throughout this section, we will sometimes make use of technical terms that we expect to be familiar to most readers. For a rigorous definition of these terms, see Chapter 2.

facie expect this ambiguity to be irreducible. As such, it is important to fully characterise and quantify this ambiguity in order to clearly understand its impact.

Another core challenge for IRL is that it must assume a specific relationship between the observed policy and the underlying reward function, i.e., it requires a specific behavioural model. In reality, the relationship between human preferences and human behaviour is incredibly complex: indeed, a complete account of this relationship would amount to a solution to many of the main questions in fields such as cognitive science, behavioural psychology, decision science, and artificial intelligence, etc. By contrast, most IRL algorithms are based on rather simple behavioural models, that typically correspond to some form of noisy optimality (c.f. Chapter 2). In fact, there are observable differences between human data and data synthesised using these standard assumptions (see, e.g., Orsini et al., 2021). This means that these behavioural models are *misspecified*, which raises the concern that they might systematically lead to flawed inferences if applied to real-world data.

Resolving the issue of misspecification in IRL is fundamentally difficult. Of course, we can incorporate findings from behavioural psychology to create behavioural models that are more and more accurate (and hence subject to less and less misspecification). Similarly, we can use machine learning to *learn* behavioural models from data (an approach pioneered by Shah, Gundotra, Abbeel, and A. Dragan, 2019), which may also yield more accurate models. However, it will never be realistically possible to create a behavioural model that is completely free from all forms of misspecification. For this reason, it is important to understand how sensitive the IRL problem is to misspecification of the underlying behavioural model: is a mostly accurate behavioural model sufficient to ensure that the inferred reward function likewise is mostly accurate, or can a slight error in the behavioural model lead to a large error in the inferred reward? In the former case misspecification may be a manageable issue, whereas in the latter case it may be practically insurmountable.

In this thesis, we provide a comprehensive theoretical study of *partial identifiability* and of *misspecification* in inverse reinforcement learning. To do this, we first introduce a cohesive theoretical framework for analysing partial identifiability and

misspecification robustness in IRL, and derive a number of core results and formal tools within this framework. We then apply these tools to exactly characterise the *ambiguity* of the reward function given several popular behavioural models, and derive necessary and sufficient conditions which exactly describe what forms of misspecification these behavioural models will tolerate. The tools we introduce can also be used to easily derive the partial identifiability and misspecification robustness of new behavioural models, beyond those we consider explicitly. Our analysis is general, as it is carried out in terms of *behavioural models*, rather than *algorithms*. This means that our results will apply to any IRL algorithm based on these behavioural models.

The motivation behind our work is to provide a theoretically principled understanding of whether and when IRL methods are (or are not) applicable to the problem of inferring a person's (true) preferences and intentions. It will never be realistically possible to fully eliminate ambiguity and misspecification from IRL, except possibly in very narrow domains. Therefore, if we wish to use IRL as a tool for preference elicitation, then it is crucial to have a good understanding of how IRL is affected by partial identifiability and misspecified behavioural models. In this thesis, we aim to contribute towards building this formal understanding.

1.2 Related Work

The issue of partial identifiability in IRL is well-known, and has been studied in a number of previous works. Indeed, the first paper to formally introduce the IRL problem (Ng and Russell, 2000) acknowledges the issue of partial identifiability, and characterises the ambiguity of the reward function under the assumption that the observed policy is optimal and the assumption that the reward of a transition $\langle s, a, s' \rangle$ only depends on the state s. This work is extended by Dvijotham and Todorov (2010), who study partial identifiability in IRL for a particular type of environment called linearly-solvable Markov decision processes (LMDPs). Partial identifiability in IRL is also studied by Cao, Cohen, and Szpruch (2021). In this paper, it is assumed that the observed policy maximises causal entropy (c.f. Chapter 2), and that the

reward of a transition $\langle s, a, s' \rangle$ only depends on the state s and action a (but not the subsequent state s'). Cao, Cohen, and Szpruch (2021) also show that the ambiguity of the reward function in this setting can be reduced by combining information from multiple environments. Also relevant is Metelli, Lazzati, and Restelli (2023), who generalise the results of Cao, Cohen, and Szpruch (2021) by also considering environments with constraints, as well as other types of regularisation. They also provide an analysis of the sample complexity of the IRL problem in this setting.

We extend this previous work on partial identifiability in IRL by providing a more complete analysis, and by integrating our analysis into the study of misspecification robustness. In particular, our analysis explicitly considers three types of policies — optimal policies, maximal causal entropy policies, and Boltzmann-rational policies. Of these, only the first two have been considered by previous works. Moreover, unlike Ng and Russell (2000) and Cao, Cohen, and Szpruch (2021), we allow the reward of a transition $\langle s, a, s' \rangle$ to depend on each of s, a, and s', and show that this reveals important additional structure that is not captured by the analysis of Ng and Russell (2000) or Cao, Cohen, and Szpruch (2021). In addition to this, we provide a general, unified framework for reasoning about both partial identifiability and misspecification robustness, and integrate our results into this framework. However, unlike Dvijotham and Todorov (2010), we will not consider LMDPs. Moreover, unlike Metelli, Lazzati, and Restelli (2023), we will not consider environments with constraints, other types of regularisation, or finite-sample bounds. Extending our analysis to cover these cases will be a direction for future work.

It is well-known that the standard behavioural models of IRL are misspecified in most applications. However, there has nonetheless so far not been much research on how sensitive IRL is to misspecification, and what forms of misspecification it can tolerate. There are previous papers which aim to *reduce* misspecification in IRL, by creating more realistic behavioural models. For example, most work in IRL assumes that the observed agent discounts exponentially. However, there is an extensive body of work in the behavioural sciences which suggests that humans are better modelled as discounting *hyperbolically* (see e.g. Thaler, 1981; Mazur, 1987; Green and Myerson,

1996; Kirby, 1997; Frederick, Loewenstein, and O'Donoghue, 2002). For this reason, Evans, Stuhlmueller, and Goodman (2015) analyse IRL for agents that use an approximation of hyperbolic discounting. Similarly, most work in IRL assumes that the observed agent is *risk-neutral*, whereas humans often are *risk-sensitive* (see e.g. Allais, 1953; Ellsberg, 1961; Kahneman and Tversky, 1979). For this reason, S. Singh et al. (2018) analyse IRL for agents with different forms of risk-sensitivity. Also relevant is Chan, Critch, and A. Dragan (2021), who provide an empirical study of IRL which incorporates many different models from the behavioural psychology literature. Chan, Critch, and A. Dragan (2021) also empirically confirm that misspecified behavioural models can lead to large errors in the inferred reward, but that this error can be reduced when the misspecification is reduced.

These approaches to reducing misspecification rely on creating more accurate behavioural models by manually incorporating more information about human behaviour. Another approach to reducing misspecification is to try to *learn* a behavioural model from data. Shah, Gundotra, Abbeel, and A. D. Dragan (2019) carry out an empirical analysis of IRL where the behavioural model and the underlying reward function are learnt in two different steps, but conclude that this approach comes with significant practical challenges. By contrast, Armstrong and Mindermann (2019) carry out a theoretical analysis of the setting where the reward function and the behavioural model are learnt at the same time, from a single stream of data. Notably, Armstrong and Mindermann (2019) derive several impossibility theorems for this setting. In particular, they show that this problem setting always will admit several degenerate solutions that fail to solve the problem in a satisfactory way, given that the learning algorithm has an inductive bias towards joint simplicity.

These earlier works all aim to *reduce* misspecification in IRL, by creating more accurate behavioural models. By contrast, we are not focusing on the problem of *reducing* misspecification. Rather, our work aims to understand how *sensitive* IRL is to misspecification of the behavioural model. As such, our analysis of misspecification is distinct from this earlier work, although it is very relevant to it. Our work aims to answer whether or not IRL will yield accurate inferences given

that the behavioural model is misspecified, which in turn would tell us how much misspecification has to be removed (be that manually or by a learning algorithm) before we can make accurate inferences about the reward function through IRL.

There are some previous papers that (like our work) study the question of how robust IRL is to misspecification of the behavioural model. In particular, Freedman, Shah, and A. Dragan (2020) study the effects of *choice set misspecification* in IRL (and reward inference more broadly), following the formalism of Jeon, Milli, and A. Dragan (2020). They also show that choice set misspecification in some cases can be catastrophic. Also relevant is Viano et al., 2021, who study the effects of misspecified *environment dynamics*. They also propose a bespoke IRL algorithm that is meant to be more robust to such misspecification. By contrast, we present a broader analysis that covers *all* forms of misspecification, within a single framework. Our work is therefore much wider in scope, and aims to provide necessary and sufficient conditions which fully describe all kinds of misspecification to which each behavioural model is robust.

Another relevant paper is J. Hong, Bhatia, and A. Dragan (2022), who also study how sensitive IRL is to misspecification of the behavioural model. Our work is more complete than this earlier work in several important respects. To start with, our problem setup is both more realistic, and more general. In particular, in order to quantify how robust IRL is to misspecification, we first need a way to formalise what it means for two reward functions to be "close". J. Hong, Bhatia, and A. Dragan (2022) formalise this in terms of the L_2-distance between the reward functions. However, this choice is problematic, because two reward functions can be very dissimilar even though they have a small L_2-distance, and vice versa (cf. Section 4.4). By contrast, our analysis is carried out in terms of specially selected *metrics* on the space of all reward functions, which are backed by strong theoretical guarantees (cf. Section 4.4). Moreover, J. Hong, Bhatia, and A. Dragan (2022) assume that there is a *unique* reward function that maximises fit to the training data, but this is violated in most real-world cases (Ng and Russell, 2000; Dvijotham and Todorov, 2010; Cao, Cohen, and Szpruch, 2021; Kim et al., 2021;

Schlaginhaufen and Kamgarpour, 2023). In addition to this, many of their results also assume "strong log-concavity", which is a rather opaque condition that is left mostly unexamined. Indeed, J. Hong, Bhatia, and A. Dragan (2022) explicitly do not answer if strong log-concavity should be expected to hold under typical circumstances. Our work is not subject to any of these limitations. Moreover, unlike J. Hong, Bhatia, and A. Dragan (2022), we also integrate our analysis of misspecification with the study of partial identifiability, which is crucial for gaining a complete understanding of the problem. In addition to this, we also present a large number of novel results that are not analogous to any results derived by J. Hong, Bhatia, and A. Dragan (2022). This includes — among other things — necessary and sufficient conditions that fully describe what kinds of misspecification many behavioural models will (or will not) tolerate.

In our analysis, we will provide a method for quantifying the difference between reward functions. Previous works have also considered this problem. In particular, Gleave et al. (2021) provide a pseudometric on the space of all reward functions, which they call EPIC (Equivalent Policy Invariant Comparison). They also show that EPIC induces a regret bound. Similarly, Wulfe et al. (2022) also provide a pseudometric for reward functions, which they call DARD (Dynamics-Aware Reward Distance). While EPIC is invariant to the transition dynamics of the environment, DARD incorporates some information about the transition function, which can lead to tighter correlation to worst-case regret. However, unlike Gleave et al. (2021), Wulfe et al. (2022) do not show that DARD induces a bound on worst-case regret. We will also introduce a family of pseudometrics on the space of all reward functions. However, unlike EPIC and DARD, our pseudometrics induce much stronger theoretical guarantees.

In our analysis, we will also provide necessary and sufficient conditions that describe when two reward functions have the same optimal policies, or the same ordering of policies. Previous work has also sought to identify transformations that can be applied to a reward function without changing some of its properties. Notably, Ng, Harada, and Russell (1999) show that if two reward functions differ by *potential*

shaping, then those reward functions have the same optimal policies. They also show that potential shaping transformations are the only *additive* transformations that have this property for any choice of transition function. We are not only concerned with additive transformations, and we are interested in reward transformations that may depend on the underlying transition function. For these reasons, we obtain results that are somewhat different from those derived by Ng, Harada, and Russell (1999). Nonetheless, potential shaping transformations will show up in this work, and will feature in many of our results.

A number of other previous works have also considered different cases in which two related Markov decision processes (MDPs) are guaranteed to have the same optimal policies in the context of *abstraction*, where the goal is to translate an MDP into an "abstract" MDP, such that the abstract MDP is easier to solve, and such that a policy that is optimal in the abstract MDP is guaranteed or likely to be exactly or approximately optimal in the original MDP (e.g. Mccallum and Ballard, 1996; Sutton, Precup, and S. Singh, 1999; Li, Walsh, and M. Littman, 2006; Givan, Dean, and Greig, 2003; Ravindran and A. Barto, 2003; Jong and Stone, 2005; Abel, Hershkowitz, and M. L. Littman, 2017; Jinnai, Abel, et al., 2019; Jinnai, Park, et al., 2019; Abel, Umbanhowar, et al., 2020; Abel, 2022). The abstracted MDP is often smaller than the original MDP (e.g. Li, Walsh, and M. Littman, 2006), but it may also be larger (e.g. Sutton, Precup, and S. Singh, 1999). In cases where the MDP is made smaller by merging certain states or actions, the transformation of the MDP into an "abstract" MDP also corresponds to a transformation of the reward function. For example, Li, Walsh, and M. Littman (2006) consider abstractions where some states in the MDP are merged, and the transition function and reward function are replaced with a weighted average of the original reward and transition function (to ensure that they are well-defined relative to the new state space). They also enumerate several different methods for deciding what states to merge, and show that some of these methods guarantee that any policy that is optimal in the new, abstract MDP also is optimal in the original MDP. Their results can easily be extended to show that the new reward function also has the same optimal

policies as the original reward function relative to the original state space and transition function, which relates their results to some of the results we will present in this work. However, unlike Li, Walsh, and M. Littman (2006), we are concerned with finding *necessary and sufficient* conditions that characterise when two reward functions have the same optimal policies, whereas Li, Walsh, and M. Littman (2006) only (indirectly) derive *sufficient* conditions. For example, the transformations derived by Li, Walsh, and M. Littman (2006) all preserve the optimal $Q^\star$-values of the reward function, but there are reward functions that have the same optimal policies but different optimal $Q^\star$-values. Moreover, not all transformations of the reward function that preserve optimal policies can be constructed by averaging the reward across certain states or actions (as already demonstrated by Ng, Harada, and Russell, 1999). Other works on creating abstract MDPs only require that the policies which are optimal in the abstract MDP are *approximately* optimal in the original MDP (e.g. Abel, Hershkowitz, and M. L. Littman, 2017; Abel, Umbanhowar, et al., 2020). Such work can be used to derive *sufficient* conditions for two reward functions to be "similar", in the sense that there is a bound on the regret that may be incurred under one reward function if the other reward function is optimised. Such results are also related to some of the results we will present in this work. However, as before, we will be concerned with finding conditions that are sufficient *and necessary*, rather than merely sufficient, and not all reward functions that are "similar" in this sense can be formed by abstracting the MDP. Other previous work on abstracting MDPs will relate to our results in similar ways.

There is also other work that studies the question of what happens if a reward function is changed or misspecified. For example, Zhuang and Hadfield-Menell (2020) consider the case when a reward function R_2 depends on a strict subset of the features which are relevant to another reward function R_1, and show that optimising R_2 in this case may lead to a policy that is arbitrarily bad according to R_1, given certain assumptions. Related to this work is also e.g. Pan, Bhatia, and Steinhardt (2022) and Pang et al. (2023), who carry out an empirical investigation of the consequences of misspecified reward functions in certain environments. Another relevant paper is

Karwowski et al. (2023), who study the effects of reward misspecification through the lens of *Goodhart's Law*, and Skalse and Abate (2023b), who provide examples of natural preference structures which cannot be expressed by reward functions at all. These papers are not concerned with deriving general methods for quantifying the difference between reward functions, nor do they characterise necessary and sufficient conditions for two reward functions to be equivalent. Our results thus complement these earlier results, but do not overlap with them.

Another body of relevant existing work is the work on the *imitation gap* in behavioural cloning and imitation learning. In this problem setting, the goal is (typically) to learn a policy that imitates the behaviour of an expert demonstrator. The imitation gap refers to the issue that if the demonstrator has access to information that is not available to the imitating agent, but a standard imitation learning algorithm is applied naïvely, then the resulting policy may have poor performance (e.g. Swamy et al., 2021; Weihs et al., 2021; Cai et al., 2022; Vuorio, Haan, et al., 2024; Vuorio, Fellows, et al., 2024). Intuitively, the imitator will learn to randomise its behaviour in states where the demonstrator would act based on information that is not available to the imitator, whereas it may be more optimal for the imitator to instead act conservatively or perform actions that let it gather more information, even if the demonstrator would not take such actions. The imitation gap can also refer to the converse issue, where the imitator has access to more information than the demonstrator. Several approaches to mitigating this issue have been proposed, some of which involve learning a reward function via IRL (e.g. Vuorio, Fellows, et al., 2024), and some which do not (e.g. Vuorio, Haan, et al., 2024). Our work is distinct from this literature in several ways. First of all, while the imitation gap can be viewed as stemming from the use of misspecified statistical models, this misspecification is best modelled as concerning the *state space* of the environment, and affects the process by which the demonstrator policy is inferred from samples from that policy. By contrast, we take both the state space of the environment and the demonstrator policy as given, and instead focus on misspecification of the model that describes how the demonstrator policy is

generated from the underlying reward function. Moreover, we aim to characterise all forms of misspecification that this model will tolerate, instead of focusing on one specific form of misspecification. Finally, the literature on the imitation gap is typically concerned with mitigating or removing this misspecification, whereas we only focus on understanding the effects of misspecification.

The literature on *third-person imitation learning* (e.g. Stadie, Abbeel, and Sutskever, 2019; Sharma, Pathak, and Gupta, 2019; Klein et al., 2023) is also tangentially related to our work. This literature concerns imitation learning where the expert demonstrations are provided from a "third-person" perspective, rather than a "first-person" perspective (or, more concretely, where the format of the observed training data does not match the required format of the learnt policy). For example, the learning algorithm may be required to solve a task by controlling a robot, but the training data is provided in the form of videos of humans solving that task, rather than in the form of input-output data for the robot. However, this literature is not primarily concerned with studying partial identifiability or misspecification, but rather, with inferring certain latent variables from a set of indirect observations. As such, its connection to our work is less direct.

There is also a range of work in (Bayesian and frequentist) statistics that studies the issues of partial identifiability and misspecification in general (but without focusing on IRL specifically). Manski (2003) is a central work on partial identifiability; this book focuses on statistical problems where a parameter cannot be identified uniquely – even in the limit of infinite data – but where it can instead be determined to lie in some *identified set* (or "identification region"). The book considers several such sampling processes, and determines what may be inferred about different partially identifiable population parameters based on these sampling processes, and how the set-valued identification regions are affected if different assumptions are imposed. Our work on partial identifiability follows a similar methodology as Manski (2003), in that we will characterise which sets of reward functions produce identical policies under a given behavioural model (meaning that those reward functions cannot be distinguished by samples from

those policies, even in the limit of infinite data). However, Manski (2003) does not specifically consider the sampling processes that are used in IRL, and so our results extend those presented by Manski (2003). Another relevant work is Imbens and Manski (2003); this work considers partially identifiable real-valued parameters, and introduces interval estimates that asymptotically cover the identification region with fixed probability. Unlike Imbens and Manski (2003), we are working with reward functions rather than real-valued parameters.[2] Moreover, we focus on infinite-data bounds, rather than finite-data bounds. Other noteworthy works include Haile and Tamer (2003) and Chernozhukov, H. Hong, and Tamer (2007), who study partial identifiability in a variety of game-theoretic models (including e.g. auction models and asset pricing models). Moon and Schorfheide (2009) compare frequentist confidence sets and Bayesian credible sets in partially identifiable models. They note that Bayesian credible sets can concentrate inside the identification region, whereas frequentist confidence sets typically extend beyond the identification region. Intuitively, this is simply because the prior of a Bayesian method may give a very low prior probability to some elements of the identification region (and for a partially identifiable model, the prior is not "washed away" in the limit of infinite data). The Bayesian setting is also studied by e.g. Liao and Simoni (2019), who focus on partially identifiable models for which the identified set is convex with a smooth boundary, whose support function is locally smooth with respect to the data distribution, and show that this setting admits a number of computationally attractive algorithms. Again, these works do not consider the IRL problem, which means that our results extend these earlier results.

In Bayesian statistics, it is common to model an inference problem as consisting of an input space $X = \mathbb{R}^n$, an output space $Y = \mathbb{R}^m$, and a joint distribution Q

[2]In particular, Imbens and Manski (2003) are concerned with the problem of inferring a single real-valued parameter θ from a set of data, and seek to identify an interval that this parameter is likely to be contained inside. However, a reward function must be defined in terms of several real-valued parameters – one for each transition. Moreover, the information that is obtained from IRL is typically only sufficient to identify the underlying reward function up to a certain equivalence class, within which the reward of each individual transition is unconstrained (c.f. Chapter 5). This means that IRL does not let us infer an upper or lower bound on the reward for each transition, but rather, lets us infer dependencies between the rewards of different transitions.

over $X \times Y$, such that the marginal distribution which Q induces over X is known, but such that the conditional probabilities $Q(y \mid x)$ (where $y \in Y$ and $x \in X$) are unknown. We also assume that we have a compact set of parameters $W \subset \mathbb{R}^d$, and a statistical model p which provides a conditional distribution $p(y \mid x, w)$ for each $y \in Y, x \in X, w \in W$, such that there is some $w \in W$ for which $p(y \mid x, w) = Q(y \mid x)$ for all x and y. The task is then to infer a posterior distribution over W given a prior distribution over W, and a number of data points $\{x, y\}$ sampled i.i.d. from Q (see e.g. Watanabe, 2009; Watanabe, 2018). In the context of IRL, we may think of W as being the space of all reward functions (though we may then have to assume that the reward function is bounded or normalised, since W typically is required to be compact). X could be the space of all possible initial states in a given MDP, and Y could be the space of all trajectories (though in that case, we would have to impose the requirement that all trajectories have a bounded length, since Y typically must be finite-dimensional). Alternatively, we may think of X as being the space of all states, and Y as being the space of all actions (though in that case, we would have to relax the assumption that each $\{x, y\}$ is sampled i.i.d.). We could also think of Y as being the space of all trajectory occupancy measures, etc. The model p corresponds to the behavioural model of the IRL algorithm. Most work in Bayesian statistics focuses on statistical models which are *regular*. Such statistical models are *identifiable*, which means that $p(y \mid x, w_1) = p(y \mid x, w_2)$ for all x, y only if $w_1 = w_2$. They are also required to have a positive-definite Fisher information matrix for all $w \in W$; this essentially amounts to assuming that the Kullback-Leibler divergence between the data distribution and the distributions induced by p for different values of w satisfy a kind of convexity condition. The model classes studied in IRL are not regular, because they are not identifiable; two different reward functions may induce exactly the same data distribution (i.e., the same observed behaviour from the demonstrator). A model which is not regular is *strictly singular*. Bayesian statistics for singular model classes is studied in *singular learning theory*, for which the most prominent current texts are Watanabe (2009) and Watanabe (2018). This literature contains several results concerning the shape of the posterior distribution

and the likelihood function for singular models, and how they change with the size of the set of training data. It also contains results such as Bayesian generalisation bounds, etc. Our work differs from the existing work in singular learning theory in several ways. First of all, singular learning theory primarily focuses on properties of general (arbitrary) singular models, whereas we focus on results that are specific to the IRL setting (and which do not generalise to arbitrary singular learning problems). Moreover, we derive several results concerning *misspecification*, which singular learning theory typically does not study. Additionally, our work is mostly not done in a Bayesian setting, and focuses on the asymptotic behaviour of learning algorithms in the limit of infinite data (whereas singular learning theory studies Bayesian uncertainty given finite data).

White (1982) is a classical text on misspecification in a frequentist setting. This paper studies maximum-likelihood inference with misspecified likelihoods, which it introduces as the *quasi-maximum likelihood estimate* (QMLE). Specifically, we assume that we have a distribution D over some set Y, a set of parameters Θ, and a family of distributions $\{P_\theta : \theta \in \Theta\}$, but where there may not be any $\theta \in \Theta$ for which $P_\theta = D$. For example, we might have that $\theta = \mathbb{R}^2$ and that P_θ is a Gaussian distribution with mean and variance θ, but that D is an exponential distribution rather than a Gaussian distribution, etc. We assume that we have a data set $\{y_i\}_{i=1}^n$ sampled from D, and that we find a θ that minimises the empirical Kullback-Leibler divergence $-\Sigma_{i=1}^n \log \mathbb{P}_{Y \sim P_\theta}(Y = y_i)$. White (1982) derives conditions under which the QMLE satisfies consistency and asymptotic normality (even when the likelihoods are misspecified), and introduces tests for detecting misspecification. In general, the conditions derived by White (1982) are not directly applicable to the problem setting studied in IRL. For example, White (1982) assumes that θ is identifiable, whereas the IRL problem typically is partially identifiable. Moreover, White (1982) is generally interested in conditions under which the estimate of θ converges (in probability) to the value of θ that minimises the (true) Kullback-Leibler divergence between D and P_θ (in other texts, this is sometimes referred to as the "pseudo-true" value of θ). However, in the IRL problem setting, we do in general not merely

want to infer a reward function that makes the observed data distribution likely (under a given behavioural model), but rather, we want to infer a reward function that is qualitatively similar to the true reward function used by the demonstrator. As such, there is a need for an analysis of misspecification that is tailored to the IRL context, which we provide in this work. The results in White (1982) are also extended and elaborated upon in works such as White (1994), etc.

Misspecification has also been studied in the Bayesian setting. Kleijn and Vaart (2012) extend the Bernstein-Von-Mises theorem to misspecified Bayesian statistical models. Formally, let $\{P_\theta : \theta \in \mathbb{R}^k\}$ be a family of distributions over a set $\mathcal{Y}$, let π be a prior distribution over $\mathbb{R}^k$, and let $\{y_i\}_{i=1}^n$ be a number of samples drawn i.i.d. from a distribution D over $\mathcal{Y}$. The standard Bernstein-Von-Mises theorem then says that the posterior distribution $\pi(\theta \mid y_1 \dots y_n)$ converges to a Gaussian distribution whose mean is the maximum likelihood estimate $\max_{\theta \in \mathbb{R}^k} \mathbb{P}_{Y_1 \dots Y_n \sim P_\theta}(Y_1 \dots Y_n = y_1 \dots y_n)$ and whose covariance matrix is given by the inverse of the Fisher Information matrix, given the assumption that the model is correctly specified (i.e., $D = P_\theta$ for some θ), the assumption that θ is identifiable (i.e., $P_{\theta_1} \neq P_{\theta_2}$ if $\theta_1 \neq \theta_2$), the assumption that the model is non-singular, and the assumption that P_θ is sufficiently smooth in θ. Kleijn and Vaart (2012) extend this theorem by relaxing the assumption that $D = P_\theta$ for some θ, and show that the posterior distribution of θ still shrinks to the point that minimises the Kullback-Leibler divergence between P_θ and D (i.e., the pseudo-true value of θ). However, they also show that Bayesian credible sets are not valid confidence sets if the model is misspecified (unlike when the model is correctly specified). As for the results derived by White (1982), the results derived by Kleijn and Vaart (2012) are not directly applicable to the IRL problem setting; Kleijn and Vaart (2012) assume that θ is identifiable (which is not the case in IRL), and consider conditions under which the estimate of θ converges to the pseudo-true value of θ (which is not the goal in IRL).

Another relevant paper on misspecification in a Bayesian setting is Müller (2013), which builds on Kleijn and Vaart (2012). Müller (2013) focuses on the "sandwich covariance matrix", which is a correction to the standard covariance

matrix used in maximum likelihood estimation. The paper shows how the sandwich covariance matrix can be used to correct the posterior distribution of a misspecified Bayesian model, and that this correction leads to asymptotically uniformly lower risk (i.e., expected loss). Frazier, Kohn, et al. (2023) is similar to Müller (2013), in that Frazier, Kohn, et al. (2023) also focuses on how to correct the posterior distribution of a misspecified Bayesian model. Their method, which they call the Q-posterior, leads to more reliable uncertainty quantification, and is applicable to a wide range of statistical models.

There is also prior work on misspecification in Bayesian statistics which focuses on more specific types of models. For example, Grünwald and Ommen (2017) study Bayesian linear regression models which assume homoskedastic data (i.e., data with noise of constant variance), but which are applied to heteroskedastic data (i.e., data where the noise may depend on the input). They show that the linear models may fail to be consistent in this setting (i.e., the parameter estimates may not converge to the pseudo-true values). They also propose a novel method, *SafeBayes*, which fixes this problem. Yang and Zhu (2018) study Bayesian methods applied to the problem of inferring phylogenetic trees. In this setting, Bayesian methods can sometimes give extremely high confidence to a model, even when this does not appear to be intuitively justified by the data, or oscillate between giving near-1 posterior probability to different models as additional data is provided. Yang and Zhu (2018) show that this behaviour is caused by misspecification in the statistical model, and provide practical recommendations for how to handle these issues in phylogenetic studies. Frazier, Robert, and Rousseau (2020) focus on approximate Bayesian computation (ABC), which is a method for approximating posterior distributions by comparing simulated data to real data, and study the case when the data simulator is misspecified. They show that the accept–reject ABC approach concentrates posterior mass around a pseudo-true parameter value, but that it may lead to incorrect credible sets (which mirrors the results derived by Kleijn and Vaart, 2012). They also show that local regression adjustment of ABC may lead to very different (and less reliable) asymptotic behaviour under model misspecification, and propose

a few new methods for detecting and diagnosing model misspecification in ABC. Unlike these papers, our work is focused on the IRL problem setting.

1.3 Contributions

This thesis makes several core contributions to the machine learning literature. First of all, in Chapter 3, we introduce a number of precise definitions that formalise what it means for an application to tolerate the ambiguity of a reward learning method, and what it means for a behavioural model to be robust to a given form of misspecification. Together, these definitions constitute a unified framework for reasoning about partial identifiability and misspecification in IRL. We also derive a number of lemmas and general results about this framework, that make it easy to reason about partial identifiability and misspecification.

In Chapter 4, we provide several results related to the issue of comparing different reward functions. Specifically, we provide necessary and sufficient conditions that describe when two reward functions have the same optimal policies, or ordering of policies. We also introduce a family of pseudometrics for continuously quantifying the difference between reward functions. We show that these pseudometrics induce both an upper and a lower bound on worst-case regret, and that any pseudometric with this property must be bilipschitz equivalent to ours.

In Chapter 5, we fully characterise the ambiguity of the reward function given several different behavioural models, and we describe the practical consequences of this ambiguity. Notably, we show that this ambiguity usually is unproblematic, but that it is too great to guarantee robust transfer to new environments.

In Chapters 6 and 7 we analyse the question of misspecification, and derive necessary and sufficient conditions that fully describe what forms of misspecification each of the standard behavioural models will tolerate. We also study a few specific types of misspecification in greater depth, such as misspecification of the parameters of the behavioural model or perturbations of the observed policy. We find that the standard behavioural models do tolerate some forms of misspecification, but that they are highly sensitive to other forms of misspecification. Notably, we find

that even mild misspecification of the discount factor γ or transition function τ can lead to very large errors in the inferred reward function.

In Chapter 8, we discuss how to extend our analysis further, by generalising the definitions we introduce in Chapter 3. Notably, we study the effects of incorporating prior knowledge about the underlying true reward function, or about the inductive bias of the reward learning algorithm, using a number of different methods. We show that most of our analysis from Chapters 4-7 carries over without any major change if our definitions are generalised.

Most of the material in this dissertation has also appeared in a number of earlier research publications, namely Skalse, Howe, et al. (2022), Skalse, Farrugia-Roberts, et al. (2022), Skalse and Abate (2023a), Skalse, Farnik, et al. (2023), and Skalse and Abate (2024). Specifically, Section 4.2 is based on some of the material from Skalse and Abate (2023a), Section 4.3 is based on Skalse, Howe, et al. (2022), and Sections 4.4-4.6 are based on Skalse, Farnik, et al. (2023). Chapter 5 and parts of Section 4.1 are based on work from Skalse, Farrugia-Roberts, et al. (2022), Chapter 6 is in large part based on the results in Skalse and Abate (2023a), and Chapter 7 is based on the results in Skalse and Abate (2024). However, this thesis also contains a number of results that cannot be found in any earlier work, especially in Section 6.2, Chapter 8, and part of Chapter 5, but also strewn throughout other sections.

Instead of trying to produce a programme to simulate the adult mind, why not rather try to produce one which simulates the child's? If this were then subjected to an appropriate course of education one would obtain the adult brain.

— Alan Turing, 1950.

2 Technical Background

In this chapter, we introduce the technical prerequisites that are needed to understand the rest of this thesis, together with our choice of notation. We also introduce all the assumptions we will make about the environment. For a more in-depth overview of reinforcement learning, see e.g. Sutton and A. G. Barto (2018), and for a more in-depth overview of inverse reinforcement learning, see e.g. Arora and Doshi (2020) or Adams, Cody, and Beling (2022).

2.1 Reinforcement Learning

A *Markov Decision Processes* (MDP) is a tuple $(\mathcal{S}, \mathcal{A}, \tau, \mu_0, R, \gamma)$ where $\mathcal{S}$ is a set of *states*, $\mathcal{A}$ is a set of *actions*, $\tau : \mathcal{S} \times \mathcal{A} \rightsquigarrow \mathcal{S}$ is a *transition function*, $\mu_0 \in \Delta(\mathcal{S})$ is an *initial state distribution*, $R : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ is a *reward function*, and $\gamma \in (0, 1)$ is a *discount rate*.[1] Here $f : X \rightsquigarrow Y$ denotes a probabilistic mapping from X to Y. A (stationary) *policy* is a function $\pi : \mathcal{S} \rightsquigarrow \mathcal{A}$, which encodes the behaviour of an

[1]Note that we exclude the case where γ is equal to 0 or 1 – this is primarily a theoretical convenience. If $\gamma = 0$ then an MDP is essentially no longer a reinforcement learning environment, which makes this case less relevant to our work. Moreover, the assumption that $\gamma > 0$ is used in Lemma 36 (and hence in all results which rely on Lemma 36), so it would take some work to extend our results to cover this case. If $\gamma = 1$ then several important quantities are no longer well-defined in general (such as Q-functions, value functions, and occupancy measures, etc), unless we make additional assumptions about the environment or the reward function. Extending our results to cover this case may be an interesting direction for future work.

agent in each state of an MDP. We use Π to denote the set of all stationary policies. A triple $\langle s, a, s' \rangle \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$ is a *transition*, and a *trajectory* $\xi = \langle s_0, a_0, s_1, a_1 \dots \rangle$ is an infinite (potentially repeating) path through an MDP, i.e. an element of $(\mathcal{S} \times \mathcal{A})^\omega$. If $s_0 \in \text{supp}(\mu_0)$ and $s_{t+1} \in \text{supp}(\tau(s_t, a_t))$ for each $t \in \mathbb{N}$, then we say that ξ is a *possible* trajectory, and otherwise it is *impossible*.

In this work, we assume that $\mathcal{S}$ and $\mathcal{A}$ are finite. Moreover, we also assume that all states in S are reachable under τ and μ_0 (i.e., for all states s, there exists a possible trajectory which includes s). This is primarily a theoretical convenience. Also note that if an MDP has unreachable states, then we may simply remove these states from $\mathcal{S}$.

The *return function* $G : (\mathcal{S} \times \mathcal{A})^\omega \to \mathbb{R}$ gives the cumulative discounted reward of each trajectory, i.e. $G(\xi) = \sum_{t=0}^{\infty} \gamma^t R(s_t, a_t, s_{t+1})$. Similarly, the *evaluation function* $J : \Pi \to \mathbb{R}$ gives the expected trajectory return of each policy, $J(\pi) = \mathbb{E}_{\xi \sim \pi}[G(\xi)]$. The *value function* $V^\pi : \mathcal{S} \to \mathbb{R}$ of a policy π encodes the expected future cumulative discounted reward from each state when following that policy π. The Q-function $Q^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ of a policy π is given by $Q^\pi(s, a) = \mathbb{E}_{S' \sim \tau(s,a)}[R(s, a, S') + \gamma V^\pi(S')]$, i.e. the expected future cumulative discounted reward conditional on taking action a in state s, and then following the policy π. Similarly, the *advantage function* of π is given by $A^\pi(s, a) = Q^\pi(s, a) - V^\pi(s)$. We say that the *ordering of policies* in an MDP is the ordering on Π that is induced by J.

Both V^π and Q^π can also be defined in terms of fixed points, because they satisfy the following *Bellman equations*:

$$V^\pi(s) = \mathbb{E}_{A \sim \pi(s), S' \sim \tau(s,A)}[R(s, A, S') + \gamma V^\pi(S')], \tag{2.1}$$

$$Q^\pi(s, a) = \mathbb{E}_{S' \sim \tau(s,a), A' \sim \pi(S')}[R(s, a, S') + \gamma Q^\pi(S', A')]. \tag{2.2}$$

Both of these equations specify a recursion, and these recursions can be shown to be contraction maps. Thus, V^π and Q^π are the only functions which satisfy Equations 2.1 and 2.2 respectively. Similarly, we can specify a unique function

$V^\star : \mathcal{S} \to \mathbb{R}$ and a unique function $Q^\star : \mathcal{S}\times\mathcal{A} \to \mathbb{R}$ via the following two Bellman recursions:

$$V^\star(s) = \max_{a\in\mathcal{A}} \mathbb{E}_{S'\sim\tau(s,a)} \left[R(s,a,S') + \gamma V^\star(S')\right], \tag{2.3}$$

$$Q^\pi(s,a) = \mathbb{E}_{S'\sim\tau(s,a)} \left[R(s,a,S') + \gamma \max_{a\in\mathcal{A}} Q^\pi(S',a)\right], \tag{2.4}$$

We refer to $V^\star$ as the *optimal value function*, and to $Q^\star$ as the *optimal Q-function*. We can also define an *optimal advantage function* $A^\star$ as $A^\star(s,a) = Q^\star(s,a) - V^\star(s)$. Note that $A^\star$ always is non-positive.

If an action a maximises $Q^\star(s,a)$ (or, equivalently, $A^\star(s,a)$) in some state s, then we say that a is an *optimal action* in s. If a policy π only takes optimal actions with positive probability, then we say that π is an *optimal policy*. We will sometimes denote an optimal policy as $\pi^\star$. If π is optimal, then π maximises the evaluation function J. However, the converse does not hold. To see this, note that π may maximise J, even if π takes sub-optimal actions in states that π visits with probability 0. Also note that if π is optimal, then $Q^\pi = Q^\star$ and $V^\pi = V^\star$. Since Equations 2.3 and 2.4 always have a solution, there is always at least one optimal policy. Moreover, the set of all optimal policies form a convex set, given by all distributions over the optimal actions in each state.

In this paper, we will often talk about pairs or sets of reward functions. In these cases, we will give each reward function a subscript R_i, and use J_i, $V_i^\star$, and V_i^π, and so on, to denote R_i's evaluation function, optimal value function, and π value function, and so on. We reserve R_0 for the reward function that is zero everywhere, i.e. $R_0(s,a,s') = 0$ for all s, a, s'. Moreover, if a reward function R satisfies that $J(\pi_1) = J(\pi_2)$ for all policies π_1, π_2, then we say that R is *trivial*. R_0 is trivial, but there are other trivial reward functions as well (c.f. e.g. Proposition 29 or Theorem 40). We will also use $\mathcal{R}$ to denote the set of all possible reward functions.

Note that we have have defined reward functions as having the type signature $\mathcal{S}\times\mathcal{A}\times\mathcal{S} \to \mathbb{R}$. In practice, it is common to instead consider reward functions with

the type signature $\mathcal{S}\times\mathcal{A}\to\mathbb{R}$. The reason for this is that, for any reward function $R_1 : \mathcal{S}\times\mathcal{A}\times\mathcal{S}\to\mathbb{R}$, we can define a second reward function $R_2 : \mathcal{S}\times\mathcal{A}\to\mathbb{R}$ as

$$R_2(s,a) = \mathbb{E}_{S'\sim\tau(s,a)}\left[R_1(s,a,S')\right].$$

It is now easy to see that $J_2 = J_1$, $Q_2^\star = Q_1^\star$, and so on. Thus, we arguably do not gain any expressive power from allowing the reward of a transition $\langle s,a,s'\rangle$ to depend on s'. However, it is important to note that R_1 and R_2 only are equivalent relative to one particular transition function τ. Moreover, in some of our results, we will quantify over multiple transition functions. We must therefore allow the reward to depend on s', to ensure that our results are fully general.

The *occupancy measure* η^π of a policy π is the $(|\mathcal{S}||\mathcal{A}||\mathcal{S}|)$-dimensional vector in which the value of the (s,a,s')'th dimension is given by

$$\sum_{t=0}^{\infty}\gamma^t\mathbb{P}_{\xi\sim\pi}\left(S_t,A_t,S_{t+1}=s,a,s'\right),$$

where the probability is over a trajectory ξ sampled from π (assuming the first state is sampled from μ_0 and transitions are sampled from τ). In other words, the occupancy measure η^π of π measures the cumulative discounted probability with which π visits each transition. To prove our results, it will sometimes be useful to map policies to their occupancy measures. One reason for this is that, if we represent the reward function R as an $(|\mathcal{S}||\mathcal{A}||\mathcal{S}|)$-dimensional vector, then $J(\pi) = \eta^\pi\cdot R$. In other words, occupancy measures allow us to decompose J into two separate steps, the first of which is independent of the reward function, and the second of which is linear. We will sometimes use Ω to denote the set of all occupancy measures, i.e. $\Omega = \{\eta^\pi : \pi\in\Pi\}$, where Π is the set of all (stationary) policies.

2.2 Inverse Reinforcement Learning

The aim of an IRL algorithm is to infer a representation of an agent's *preferences* based on their *behaviour*. It is typically assumed that these preferences can be represented as a reward function, and that the observed behaviour has the form of a policy. It is also typically assumed that the environment of the agent can be

modelled as an MDP. The IRL problem can thus loosely be stated as follows. There is an unknown reward function R. You get to observe a policy π, which has been computed from R relative to some transition function τ, initial state distribution μ_0, and discount factor γ. We may or may not assume that τ, μ_0, and γ are known.[2] The goal is then to infer a reward function R_H, that is as similar as possible (in some relevant sense) to the true reward function R.

An IRL algorithm must make assumptions about how the observed policy π relates to the underlying reward function, R. These assumptions are referred to as the *behavioural model*. In some cases, the behavioural model simply assumes that π is optimal under R (e.g. Ng and Russell, 2000). However, this assumption is often unrealistic; people sometimes make mistakes, and are subject to limited information and limited cognitive resources. As such, many IRL algorithms make use of other behavioural models. One common model is *Boltzmann rationality* (e.g. Ramachandran and Amir, 2007), which says that

$$\mathbb{P}(\pi(s) = a) = \left(\frac{\exp \beta Q^\star(s,a)}{\sum_{a' \in \mathcal{A}} \exp \beta Q^\star(s,a')} \right).$$

Here $\beta \in \mathbb{R}^+$ is known as a *temperature parameter*. When π satisfies this relationship, we refer to it as a *Boltzmann-rational policy*. The function $f : \mathbb{R}^n \to \mathbb{R}^n$ given by $f(v)_i = \exp \beta v_i / \sum_{j=1}^n \exp \beta v_j$ is known as the *softmax function* for temperature β. Thus, a Boltzmann-rational policy is given by applying a softmax function to the optimal Q-function. Intuitively speaking, such a policy takes every action with positive probability, but is more likely to take actions with high value than actions with low value. Boltzmann-rationality can therefore be seen as a form of noisy optimality.

Another common behavioural model is *causal entropy maximisation* (e.g. Ziebart, 2010). This behavioural model specifies an alternative optimisation criterion, known as the *maximal causal entropy* (MCE) objective:

$$J^{\mathrm{MCE}}(\pi) = \mathbb{E}_{\xi \sim \pi} \left[\sum_{t=0}^{\infty} \gamma^t (R(s_t, a_t, s_{t+1}) + \alpha H(\pi(s_t))) \right].$$

[2]Note that we generally have to assume that the set of states $\mathcal{S}$ and the set of actions $\mathcal{A}$ are known, since these are the domain and codomain of π.

Here $\alpha \in \mathbb{R}^+$ is a *weight*, and H is the Shannon entropy function. A policy π which maximises the MCE objective is referred to as an MCE policy. Moreover, let the *soft Q-function* $Q^{\mathrm{S}}_{\alpha} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ be the function that is defined by the following Bellman recursion:

$$Q^{\mathrm{S}}_{\alpha}(s,a) = \mathbb{E}_{S' \sim \tau(s,a)} \left[R(s,a,S') + \gamma \alpha \log \sum_{a' \in \mathcal{A}} \exp\left(\left(\frac{1}{\alpha} \right) Q^{\mathrm{S}}_{\alpha}(S',a') \right) \right] \tag{2.5}$$

This recursion is a contraction map, and thus has a unique solution. Moreover, it can be shown that the MCE policy is given by

$$\mathbb{P}(\pi(s) = a) = \left(\frac{\exp(1/\alpha) Q^{\mathrm{S}}_{\alpha}(s,a)}{\sum_{a' \in \mathcal{A}} \exp(1/\alpha) Q^{\mathrm{S}}_{\alpha}(s,a')} \right),$$

i.e., by applying the softmax function with temperature $1/\alpha$ to Q^{S}_{α} (see Haarnoja et al., 2017, their Theorem 1 and 2). Note that this implies that the MCE policy always is unique. Intuitively speaking, the MCE policy maximises expected cumulative discounted reward, subject to a regularisation term that encourages the policy to be as stochastic as possible. One way to justify the MCE objective as a model of human behaviour is to note that a boundedly rational agent presumably is less likely to solve a given problem using a strategy that is highly sensitive to mistakes.[3]

In the current literature, most IRL algorithms assume that the observed policy is either optimal, Boltzmann-rational, or MCE optimal. Therefore, we will refer to these behavioural models as the *standard* behavioural models, and focus on them in our analysis. We will however additionally present many results that hold for wider classes of behavioural models.

There are many ways to design an IRL algorithm around a given behavioural model (see e.g. Ng and Russell, 2000; Ramachandran and Amir, 2007; Ziebart, 2010; Haarnoja et al., 2017, etc). However, the details of these algorithms will not be important for understanding this work, because our analysis will be carried

[3]As an intuitive example, suppose you are choosing between two different train routes between a point A and some destination B, where the first route is a direct connection, and the second route involves several stops. Suppose also that the second route is slightly faster if you do not miss any connecting trains, but longer if you do miss one or more of the connections. A boundedly rational agent may then be more likely to pick the first route, even if an optimal agent would pick the second route. The entropy regularisation in the MCE objective roughly captures this kind of reasoning, which may make it a plausible model of bounded rationality.

out primarily in terms of behavioural models, rather than specific algorithms. In this way, we can derive results that apply to any IRL algorithm that is based on a given behavioural model.

2.3 Metrics, Pseudometrics, and Norms

In our analysis, we will often quantify the difference between different kinds of objects (especially reward functions). To do this, we will make use of *metrics*, *pseudometrics*, and *norms*. Given a set X, a function $m : X \times X \to \mathbb{R}$ is a *pseudometric* on X if it satisfies the following axioms:

1. Indiscernibility of identicals: $m(x, x) = 0$ for all $x \in X$.
2. Positivity: $m(x, y) \geq 0$ for all $x, y \in X$.
3. Symmetry: $m(x, y) = m(y, x)$ for all $x, y \in X$.
4. Triangle inequality: $m(x, z) \leq m(x, y) + m(y, z)$ for all $x, y, z \in X$.

If m additionally satisfies the identity of indiscernibles, which says that $m(x, y) \neq 0$ for all $x, y \in X$ such that $x \neq y$, then m is a *metric*. Every metric is a pseudometric, but not vice versa.

Given a vector space V, a function $n : V \to \mathbb{R}$ is a *norm* on V if n satisfies the following axioms:

1. Non-negativity: $n(v) \geq 0$ for all $v \in V$.
2. Positive definiteness: $n(v) = 0$ if and only if v is the zero vector.
3. Absolute homogeneity: $n(c \cdot v) = c \cdot n(v)$ for all $v \in V$ and $c \in \mathbb{R}$.
4. Triangle inequality: $n(v + w) \leq n(v) + n(w)$ for all $v, w \in V$.

For any real number $p \geq 1$, the function $L_p : \mathbb{R}^d \to \mathbb{R}$ given by

$$L_p(v) = \left(\sum_{i=1}^{d} |v_i|^p \right)^{1/p}$$

is a norm on $\mathbb{R}^d$. The function $L_\infty : \mathbb{R}^d \to \mathbb{R}$, given by $L_\infty(v) = \max_i |v_i|$, is also a norm. Moreover, if $n : \mathbb{R}^d \to \mathbb{R}$ is a norm , and $M : \mathbb{R}^d \to \mathbb{R}^d$ is an invertible matrix, then $n \circ M$ is also a norm.

If $n : \mathbb{R}^n \to \mathbb{R}$ is a norm, then the function $m : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ given by $m(v, w) = n(v - w)$ is a metric on $\mathbb{R}$. For convenience, we will (in a mild overload of notation) also denote this metric using n, so that e.g.

$$L_2(v, w) = \left(\sum_{i=1}^{|v|} |v_i - w_i|^2 \right)^{1/2},$$

and so on. Every norm corresponds to a metric in this way, but not every metric corresponds to a norm.

Given a set X, and two metrics m_1, m_2 on X, if there exists positive constants $\ell, u \in \mathbb{R}^+$ such that

$$\ell \cdot m_1(x, y) \leq m_2(x, y) \leq u \cdot m_1(x, y)$$

for all $x, y \in X$, then m_1 and m_2 are said to be *bilipschitz equivalent*. All norms (but not all metrics) are bilipschitz equivalent on any finite-dimensional vector space.

The only way to rectify our reasonings is to make them as tangible as those of the Mathematicians, so that we can find our error at a glance, and when there are disputes among persons, we can simply say: Let us calculate, without further ado, to see who is right.

— Gottfried Wilhelm Leibniz, 1685.

3 New Definitions and Formalisms

In this chapter, we introduce the theoretical frameworks that underpin our further analysis. First, we will introduce a number of definitions that formalise the notion of *partial identifiability*. After this, we will introduce two related but distinct ways of formalising *misspecification robustness*, and discuss the benefits of each approach. In addition, we will present several relevant intermediate results about our framework. These lemmas will be used to prove our later results, but are also insightful in their own right.

Some of the definitions we provide in this section will be given relative to an equivalence relation $\equiv$ or a pseudometric $d^{\mathcal{R}}$ on $\mathcal{R}$, the set of all possible rewards. The purpose of these is to quantify differences between reward functions (in particular, the difference between the learnt reward function and the true reward function). In this section, we will not discuss the issue of which equivalence relation $\equiv$ or pseudometric $d^{\mathcal{R}}$ to use — this question will instead be addressed in Chapter 4.

3.1 Partial Identifiability

In this section, we describe the framework that we will use to analyse *partial identifiability*. Before going into the specifics, let us recall the details of the problem. In IRL, there are typically multiple reward functions that are consistent with a

given data source, even in the limit of infinite data. This means that the reward function is ambiguous, or *partially identifiable*, based on such data sources. We wish to characterise this ambiguity.

At the same time, it is important to note that it often is unnecessary to identify a reward function uniquely, because all plausible reward functions might lead to the same outcome in a given application. For example, if we want to learn a reward function in order to compute an optimal policy, then it is enough to learn a reward function that has the same optimal policies as the true reward function. It is therefore important to also consider the *ambiguity tolerance* of various applications.

Our framework for characterising partial identifiability in IRL is based on the following three definitions:

Definition 1. We say that a *reward object* is a function $f : \mathcal{R} \to X$, where $\mathcal{R}$ is the set of all reward functions, and X is any set. If X is the set of all policies Π, then we refer to f as a *behavioural model*.

Definition 2. Given a reward object $f : \mathcal{R} \to X$, the *invariance partition* $\text{Am}(f)$ of f is the partition of $\mathcal{R}$ according to the equivalence relation $\equiv_f$ where $R_1 \equiv_f R_2$ if and only if $f(R_1) = f(R_2)$.

Definition 3. Given two partitions P, Q of $\mathcal{R}$, if $R_1 \equiv_P R_2 \implies R_1 \equiv_Q R_2$ then we write $P \preceq Q$. Given two reward objects $f : \mathcal{R} \to X$, $g : \mathcal{R} \to Y$, if $\text{Am}(f) \preceq \text{Am}(g)$ then we say that f is *no more ambiguous* than g. If $\text{Am}(f) \preceq \text{Am}(g)$ but not $\text{Am}(g) \preceq \text{Am}(f)$, then we write $\text{Am}(f) \prec \text{Am}(g)$ and say that f is *strictly less ambiguous* than g.

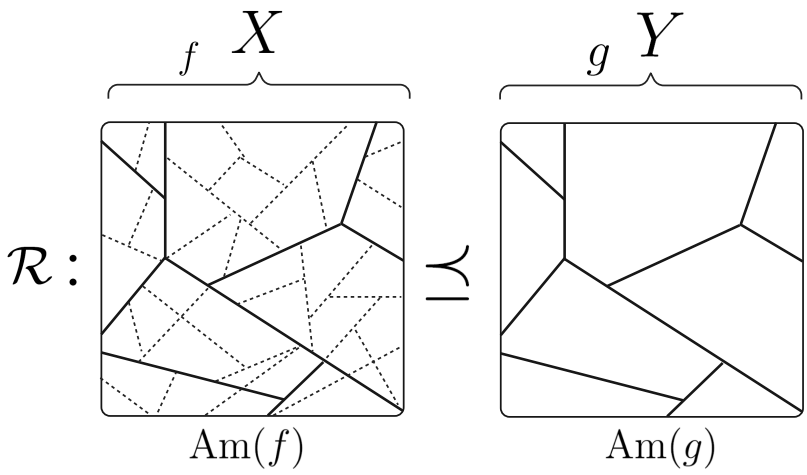


Figure 3.1: This figure illustrates Definition 1-3 visually. Specifically, suppose $f : \mathcal{R} \to X$ and $g : \mathcal{R} \to Y$ are functions (or "reward objects" in our terminology). Now f induces a partitioning $\text{Am}(f)$ of $\mathcal{R}$ according to which R_1 and R_2 belong to the same partition if (and only if) $f(R_1) = f(R_2)$, and likewise for g and $\text{Am}(g)$. If $g(R_1) = g(R_2)$ whenever $f(R_1) = f(R_2)$, then $\text{Am}(f)$ is a partition refinement of $\text{Am}(g)$, which can be visualised as in the figure above. This corresponds to the case when $\text{Am}(f) \preceq \text{Am}(g)$, where f is *less ambiguous* than g.

Before moving on, let us provide an intuitive explanation of these definitions. First of all, anything that can be computed from a reward function can be seen as a *reward object*. For example, we could consider the function $f_{\tau,\gamma}$ that, given a reward function R, returns the optimal Q-function given transition function τ and discount factor γ. In this case, X would be the set of functions $\mathcal{S}\times\mathcal{A} \to \mathbb{R}$. Similarly, we could consider the function $b_{\tau,\gamma,\beta}$ that returns the Boltzmann-rational policy for temperature β, given transition function τ and discount factor γ. In this case X would be the set Π encompassing all policies, which means that $b_{\tau,\gamma,\beta}$ is a behavioural model. As we can see, reward objects (as defined in Definition 1) are a versatile abstract building block that can be used for complex constructions. We will mainly, but not exclusively, consider reward objects with the type $\mathcal{R} \to \Pi$, i.e. behavioural models.

We can use reward objects to create an abstract model of a reward learning algorithm $\mathcal{L}$ as follows; first, we assume that there is a true underlying reward

function $R^\star$. We model the data source as a function $f : \mathcal{R} \to X$, for some data space X, so that the learning algorithm observes $f(R^\star)$. Note that $f(R^\star)$ could be a distribution, which models the case where the data comprises a set of samples from some source, but it could also be a single finite object. Next, we assume that $\mathcal{L}$ learns (or converges to) a reward function R_H that is compatible with the observed data, which means that $f(R_H) = f(R^\star)$. Note that this primarily is a model of the *asymptotic* behaviour of learning algorithms, in the limit of *infinite data*.

Hence, the defined *invariance partition* $\text{Am}(f)$ of f groups together all reward functions that a learning algorithm $\mathcal{L}$ that is based on f could converge to. For example, let $b_{\beta,\tau,\gamma} : \mathcal{R} \to \Pi$ be the function that returns the Boltzmann-rational policy for temperature β given transition function τ and discount factor γ. If two reward functions R_1, R_2 have the same Boltzmann-rational policy — i.e., if $b_{\beta,\tau,\gamma}(R_1) = b_{\beta,\tau,\gamma}(R_2)$ — then R_1 and R_2 cannot be distinguished by $b_{\beta,\tau,\gamma}$. Thus, $\text{Am}(b_{\beta,\tau,\gamma})$ partitions $\mathcal{R}$ according to which reward functions can and cannot be separated by a learning algorithm based on Boltzmann-rational policies. This means that $\text{Am}(f)$ describes the *ambiguity* of the reward R given the data $f(R)$.

Next, note that we can also interpret the invariance partition of f as a characterisation of the information that we need to have about R to construct $f(R)$. Specifically, let $g : \mathcal{R} \to Y$ be a function whose output we wish to compute. If $R^\star$ is the true reward function, then it is acceptable to instead learn a reward function R_H as long as $g(R_H) = g(R^\star)$. This means that the invariance partition of g also groups together all reward functions that it would be acceptable to learn, for the purpose of computing the output of g. Stated differently, $\text{Am}(g)$ describes the *ambiguity tolerance* of R when computing the value of $g(R)$.

We can now see that $\preceq$ formalises two important relationships between reward objects. First of all, if f and g correspond to two different reward learning data sources, and $\text{Am}(f) \prec \text{Am}(g)$, then we get strictly more information about the underlying reward function by observing data from f than we get by observing data from g. Moreover, if f is a reward learning data source and g is a downstream application, then $\text{Am}(f) \preceq \text{Am}(g)$ is precisely the condition of g tolerating the ambiguity

of the data source f (i.e., any two reward functions that cannot be distinguished by data from f lead to identical outputs when computing the value of g).

To make this more intuitive, let us discuss an example. Consider first a reward learning data source, such as trajectory comparisons. In this case, we can let X be the set of all (strict, partial) orderings of the set of all trajectories, and f be the function that returns the ordering of the trajectories that is induced by the trajectory return function, G. Let $R^\star$ be the true reward function. In the limit of infinite data, the reward learning algorithm will learn a reward function R_H that induces the same trajectory ordering as $R^\star$, which means that $f(R_H) = f(R^\star)$. Furthermore, if we want to use the learnt reward function to compute a policy, then we may consider the function $g : \mathcal{R} \to \Pi$ that takes a reward function R, and returns a policy $\pi^\star$ that is optimal under R (given some τ and γ). Then if $f(R_H) = f(R^\star) \implies g(R') = g(R^\star)$, we will compute a policy that is optimal under the true reward $R^\star$. This corresponds to the condition that $\mathrm{Am}(f) \preceq \mathrm{Am}(g)$.

Before moving on, we should also briefly add a remark on our use of the term "ambiguity" in this thesis. We will mostly use this as a general term, corresponding to a number of related technical notions in different contexts. In particular, "the ambiguity of f" and "the ambiguity of the reward function under f" should both normally be taken to refer to $\mathrm{Am}(f)$. Similarly, "f is less ambiguous than g" and "g tolerates the ambiguity of f" should be understood as "$\mathrm{Am}(f) \preceq \mathrm{Am}(g)$", as per Definition 3. If we say that f is "too ambiguous" (for some purpose), then this should normally be taken to mean that $\mathrm{Am}(f) \not\preceq P$ (for some partition P of $\mathcal{R}$). The intended meaning should generally be clear in each given context. However, to avoid any risk of confusion, all theorems and proofs are stated in terms of unambiguous technical terms.

Next, it is useful to note that the invariances of an object are inherited by all objects that can be computed from it. More formally:

Lemma 4. *Consider two reward objects $f : \mathcal{R} \to X$, $g : \mathcal{R} \to Y$. If there exists a function $h : X \to Y$ such that $h \circ f = g$, then* $\mathrm{Am}(f) \preceq \mathrm{Am}(g)$.

Proof. If $f(R_1) = f(R_2)$, then $h \circ f(R_1) = h \circ f(R_2)$, so $g(R_1) = g(R_2)$. Thus $f(R_1) = f(R_2) \implies g(R_1) = g(R_2)$, so $\mathrm{Am}(f) \preceq \mathrm{Am}(g)$. □

This simple observation has the important consequence that if there is an intermediate object (e.g., a Q-function) that is too ambiguous for a given application, then this ambiguity will also hold for any object that can be computed from this intermediate object.

Note that $\preceq$ is transitive; if $P \preceq Q$ and $Q \preceq R$ then $P \preceq R$. It is also antisymmetric; if $P \preceq Q$ and $Q \preceq P$ then $P = Q$. This means that our framework endows all reward learning data sources and applications with a lattice structure, where $f \to g$ if $\mathrm{Am}(f) \preceq \mathrm{Am}(g)$. This lattice structure enables reading out several important relationships graphically:

1. If $f \to g$ then a data source based on f is at least as informative as a data source based on g.

2. If $f \to g$ then a data source based on f contains enough information to compute the output of g.

3. If $f \to g$ then it is in principle possible to compute $g(R)$ from $f(R)$.

4. If $f \to g$ and $f \not\to h$ then $g \not\to h$. In other words, if f is a data source that does not contain enough information to compute h, and g is a data source that can be derived from f, then g does not contain enough information to compute h.

As such, our definitions make it easy to reason about partial identifiability, ambiguity, and ambiguity tolerance, within a single unified framework.

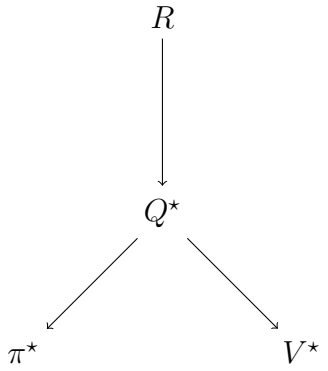


Figure 3.2: This figure gives a simple illustration of how Definition 1-3 induces a partial order over objects that can be computed from reward functions. For example, let q be the function that, given a reward function R, returns the optimal Q-function $Q^\star$, and let v be the function that, given a reward function R, returns the optimal value-function $V^\star$. Since $V^\star$ can be computed from $Q^\star$, we have that $\text{Am}(q) \preceq \text{Am}(v)$, which can be represented as $q \to v$ (or $Q^\star \to V^\star$) in a figure. Important relationships between data sources can then be read out graphically — for example, if $Q^\star$ is too ambiguous for a given application, then $V^\star$ must be too ambiguous as well.

Next, it will often be useful to express $\text{Am}(f)$ in terms of the set of all transformations of R that preserve $f(R)$. Formally:

Definition 5. A *reward transformation* is a map $t : \mathcal{R} \to \mathcal{R}$. We say that the *invariances* of f is a set of reward transformations T if for all $R_1, R_2 \in \mathcal{R}$, we have that $f(R_1) = f(R_2)$ if and only if there is a $t \in T$ such that $t(R_1) = R_2$. We then say that f *determines* R *up to* T.

Moreover, when talking about a particular kind of object, we will for the sake of brevity sometimes leave the function f implicit, and instead just mention the relevant object. For example, we might say that "the Boltzmann-rational policy determines R up to T". This should be understood as saying that "f determines R up to T, where f is the function that takes a reward and returns the corresponding Boltzmann-rational policy". It is also worth noting that f and T often will be parameterised by τ, μ_0, or γ; this dependence will usually be spelt out, but it may in some cases be omitted when it is unambiguous from the context. Moreover, we will sometimes express the invariances of a function f in terms of several sets of reward transformations – for example, we might say that "f determines R up to T_1

and T_2". This should be understood as saying that f determines R up to T, where T is the set of all transformations that can be formed by composing transformations in T_1 and T_2 (in any order). For more details, see Section 3.4.

It is worth noting that if f determines R up to T_1, and g determines R up to T_2, where $T_1 \subseteq T_2$, then $\mathrm{Am}(f) \preceq \mathrm{Am}(g)$. Similarly, if f determines R up to T_1, and g determines R up to T_1 *and* T_2, then we also have that $\mathrm{Am}(f) \preceq \mathrm{Am}(g)$. This ought to be quite intuitive, but noting this will make it easier to compare our theorem statements in Chapter 5.

Note that the notions introduced in Definition 1-3 only let us compare the ambiguity of different data sources f and g by examining whether or not the ambiguity of f is strictly greater than the ambiguity of g, or vice versa. This means that Definition 1-3 do not let us *quantify* the absolute ambiguity of a data source. To address this, we introduce the following definition:

Definition 6. Given a set of reward functions $S \subseteq \mathcal{R}$, and a pseudometric $d^{\mathcal{R}}$ on $\mathcal{R}$, we say that the *diameter* $\mathrm{diam}(S)$ of S is the supremum of the distance between pairs of reward functions in S under $d^{\mathcal{R}}$, i.e.

$$\mathrm{diam}(S) = \sup\{d^{\mathcal{R}}(R_1, R_2) : R_1, R_2 \in S\}.$$

Moreover, given a reward object $f : \mathcal{R} \to X$, we say that the *upper diameter* of $\mathrm{Am}(f)$ is the *greatest* diameter of any set in $\mathrm{Am}(f)$, i.e.

$$\sup\{\mathrm{diam}(S) : S \in \mathrm{Am}(f)\}.$$

Similarly, we say that the *lower diameter* of $\mathrm{Am}(f)$ is the *smallest* diameter of any set in $\mathrm{Am}(f)$, i.e.

$$\inf\{\mathrm{diam}(S) : S \in \mathrm{Am}(f)\}.$$

To understand these definitions, note that if we have a pseudometric $d^{\mathcal{R}}$ on $\mathcal{R}$ that provides a quantification of how different any two reward functions are, then we can use this pseudometric to measure the "size" of $\mathrm{Am}(f)$, where a larger size corresponds to greater ambiguity. Since not every set in $\mathrm{Am}(f)$ may have the same

size, we further distinguish between the upper and the lower diameter of $\mathrm{Am}(f)$. Intuitively, the upper diameter measures the worst-case ambiguity of the reward function under f, whereas the lower diameter measures the best-case ambiguity. For example, if the lower diameter of $\mathrm{Am}(f)$ is ϵ, then that means that there for *any* $x \in \mathrm{Im}(f)$ exists two reward functions R_1, R_2 such that $f(R_1) = f(R_2) = x$, but such that the distance between R_1 and R_2 is ϵ (or arbitrarily close to ϵ). By contrast, if the upper diameter of $\mathrm{Am}(f)$ is ϵ, then that means that there is *some* $x \in \mathrm{Im}(f)$ for which there exists two reward functions R_1, R_2 such that $f(R_1) = f(R_2) = x$, but such that the distance between R_1 and R_2 is ϵ (or arbitrarily close to ϵ). Also note that the upper diameter of $\mathrm{Am}(f)$ always is at least as great as the lower diameter of $\mathrm{Am}(f)$.

It is also important to note that, while the (upper and lower) diameter of $\mathrm{Am}(f)$ provides a way of quantifying the size of $\mathrm{Am}(f)$, this does not capture all of the structure of $\mathrm{Am}(f)$. For example, even if the lower diameter of $\mathrm{Am}(f)$ is greater than the upper diameter of $\mathrm{Am}(g)$, this does *not* guarantee that $\mathrm{Am}(g) \preceq \mathrm{Am}(f)$. For this reason, it may in some cases be more informative to characterise $\mathrm{Am}(f)$ in terms of reward transformations (rather than in terms of its upper and lower diameter), since this provides a complete description of the structure of $\mathrm{Am}(f)$.

3.2 Misspecification Robustness

In this section, we introduce the frameworks that we will use for analysing robustness to misspecification. Do do this, we will fist give an abstract model of a reward learning algorithm $\mathcal{L}$ that is slightly more general than that provided in Section 3.1. As before, we assume that there is a true underlying reward function $R^\star$, and that the training data is generated by a function $g : \mathcal{R} \to X$, so that the learning algorithm observes $g(R^\star)$. Moreover, we assume that the learning algorithm $\mathcal{L}$ has a model $f : \mathcal{R} \to X$ of how the observed data relates to $R^\star$, such that $\mathcal{L}$ converges to a reward function R_H that satisfies $f(R_H) = g(R^\star)$. However, unlike in Section 3.1, we will not assume that $f = g$; this allows us to reason about the impact of misspecification. If $f \neq g$, then f is *misspecified*, otherwise f is correctly specified.

Intuitively, we want to say that f is robust to misspecification with g if a learning algorithm $\mathcal{L}$ that is based on f is guaranteed to learn a reward function that is "close" to the true reward function if it is trained on data generated from g. To make this statement formal, we need a definition of what it means for two reward functions to be "close". Our first formalisation defines this in terms of *equivalence classes*. Specifically, we assume that we have a partition P of $\mathcal{R}$ (which, of course, corresponds to an equivalence relation), and that the learnt reward function R_H is "close enough" to the true reward $R^\star$ if they are in the same equivalence class, $R_H \equiv_P R^\star$. We will for the time being leave out the question of how to pick the partition P, and later revisit this question in Chapter 4. Given this, we can now provide our first definition of robustness to misspecification.

Definition 7. Given a partition P of $\mathcal{R}$, and two reward objects $f, g : \mathcal{R} \to X$, we say that f is *P-robust to misspecification* with g if each of the following conditions are satisfied:

1. If $f(R_1) = g(R_2)$ then $R_1 \equiv_P R_2$.
2. $\mathrm{Im}(g) \subseteq \mathrm{Im}(f)$.
3. $\mathrm{Am}(f) \preceq P$.
4. $f \neq g$.

Let us explain each of these conditions. The first condition says that if f is P-robust to misspecification with g, then any learning algorithm $\mathcal{L}$ based on f is guaranteed to learn a reward function that is P-equivalent to the true reward function when trained on data generated from g. This is the core property of misspecification robustness, which ensures that the mismatch between f and g is unproblematic.

The second condition ensures that $\mathcal{L}$ can never observe data that is impossible according to its assumed model. For example, suppose f maps each reward function to a deterministic policy; in that case, the learning algorithm $\mathcal{L}$ will assume that the observed policy must be deterministic. What happens if such an algorithm is given data from a nondeterministic policy? This is undefined, absent further details about

$\mathcal{L}$. Since we do not want to make any strong assumptions about $\mathcal{L}$, it is reasonable to require that any data that could be produced by g, can be explained under f.

The third condition says that any learning algorithm $\mathcal{L}$ based on f is guaranteed to learn a reward function that is P-equivalent to the true reward function when trained on data generated by f, i.e. when there is no misspecification. In other words, f is *no more ambiguous* than P, in the sense of Definition 3. This condition is included to rule out certain uninteresting edge cases — we discuss this in more detail at the end of this section. The final condition simply says that f and g are distinct — if they are not, then f is not misspecified![1]

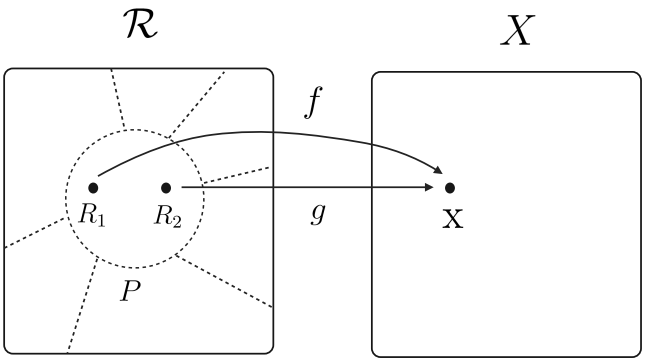


Figure 3.3: This figure illustrates the conditions in Definition 7. Both f and g are functions from the space of rewards $\mathcal{R}$ to a set X, and P is a partitioning of $\mathcal{R}$. The learning algorithm $\mathcal{L}$ observes $x = g(R_1)$ for some unknown reward function R_1, and will find a reward function R_2 such that $f(R_2) = x$. We wish to ensure that $R_2 \equiv_P R_1$. If this holds for all R_2 and R_1 such that $f(R_2) = g(R_1)$, together with the other conditions in Definition 7, when we say that f is P-robust to misspecification with g.

Our next definition formalises misspecification robustness in terms of pseudo-metrics on $\mathcal{R}$ (rather than equivalence relations). While Definition 7 captures many important properties of misspecification, it is also limited by the fact that

[1]Note that if we were to drop condition 4, and set $f = g$, then Definition 7 would be equivalent to Definition 3, which is the definition we use to formalise ambiguity tolerance. Definition 7 is thus an extension of Definition 3, designed to cover the case where the true data generating process (i.e. g) is different from model assumed by $\mathcal{L}$ (i.e. f).

it quantifies the differences between reward functions in terms of equivalence relations. With this definition, two reward functions are either equivalent or not, which means that Definition 7 cannot distinguish between small and large errors in the learned reward function. To alleviate this limitation, we introduce a second definition of misspecification robustness that is based on *pseudometrics* on $\mathcal{R}$; this will let us quantify the error in the learnt reward function in a fine-grained and continuous manner.

Definition 8. Given a pseudometric $d^{\mathcal{R}}$ on $\mathcal{R}$, and two reward objects $f, g : \mathcal{R} \to X$, we say that f is ϵ-robust to misspecification with g as measured by $d^{\mathcal{R}}$ if each of the following conditions are satisfied:

1. If $f(R_1) = g(R_2)$ then $d^{\mathcal{R}}(R_1, R_2) \leq \epsilon$.

2. $\text{Im}(g) \subseteq \text{Im}(f)$.

3. If $f(R_1) = f(R_2)$ then $d^{\mathcal{R}}(R_1, R_2) \leq \epsilon$.

4. $f \neq g$.

The conditions in Definition 8 mirror the conditions in Definition 7; the second and fourth conditions are identical in both definitions, and the first and third conditions are restated in terms of a pseudometric $d^{\mathcal{R}}$. We will for the time being leave out the question of how to pick a pseudometric $d^{\mathcal{R}}$, and later revisit this question in Chapter 4.

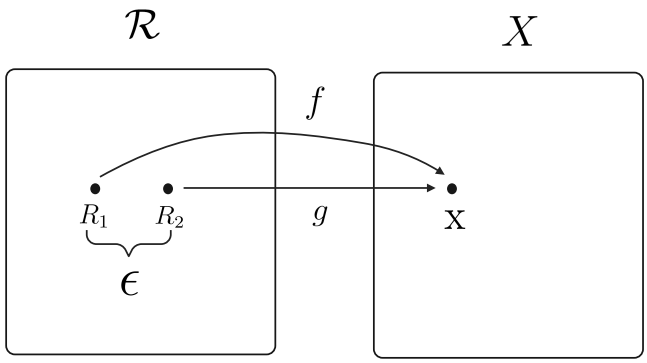


Figure 3.4: This figure illustrates the conditions in Definition 8. Both f and g are functions from the space of all rewards $\mathcal{R}$ to some set X, and we have some pseudometric $d^{\mathcal{R}}$ on $\mathcal{R}$. The learning algorithm $\mathcal{L}$ observes $x = g(R_1)$ for some unknown reward function R_1, and will find a reward function R_2 such that $f(R_2) = x$. We wish to ensure that $d^{\mathcal{R}}(R_1, R_2) \leq \epsilon$. If this holds for all R_1 and R_2 such that $f(R_2) = g(R_1)$, together with the other conditions in Definition 8, when we say that f is ϵ-robust to misspecification with g (as measured by the pseudometric $d^{\mathcal{R}}$).

Next, note that any result expressed in terms of Definition 7 can be translated into a corresponding result expressed in terms of Definition 8:

Proposition 9. *Consider a pseudometric $d^{\mathcal{R}}$ and an equivalence relation $\equiv_P$ on $\mathcal{R}$ such that $R_1 \equiv_P R_2$ if and only if $d^{\mathcal{R}}(R_1, R_2) = 0$. Then f is P-robust to misspecification with g if and only if f is 0-robust to misspecification with g as measured by $d^{\mathcal{R}}$.*

Proof. Immediate from Definition 7 and 8. □

Also note that if f is 0-robust to misspecification with g, then it of course follows that f is ϵ-robust to misspecification with g for all $\epsilon > 0$:

Proposition 10. *If f is ϵ-robust to misspecification with g measured by $d^{\mathcal{R}}$, and $\delta > \epsilon$, then f is δ-robust to misspecification with g measured by $d^{\mathcal{R}}$.*

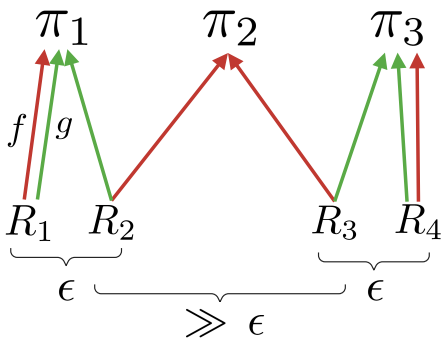
Proof. Immediate from Definition 8. □

In light of this, one might ask why we should use Definition 7 if Definition 8 is strictly more expressive. There are several reasons for this. First of all, Definition 7 still captures most of what we care about in practice, while also being notably easier to work with. Moreover, while any pseudometric can be straightforwardly translated into an equivalence relation, it is not always straightforward to translate an equivalence relation into a metric, other than by letting this metric be equal to 0 for equivalent reward functions and 1 for non-equivalent reward functions. Additionally, Definition 7 will let us derive results that are both stronger and easier to interpret qualitatively, than what is possible using Definition 8. For this reason, we will make use of both Definition 7 and Definition 8.

In Chapter 8, we provide a more extensive discussion of Definition 7 and 8, including ways in which these definitions may be modified or generalised, and whether such modifications would have a meaningful impact on any of our results. We show that many natural generalisations would lead to results that are identical or closely analogous to the results that we will provide for Definition 7 and 8.

Before moving on, let us briefly comment on the third condition in Definition 7 and Definition 8. This condition informally says that for f to be robust to misspecification with g, it is necessary that a learning algorithm which is based on f should be guaranteed to learn a reward function that is close to the true reward function when there is no misspecification. It may not be immediately obvious why this assumption is included, since we assume that the data is generated by g, where $f \neq g$. Let us therefore provide an example, to explain the motivation for this condition.

Let R_1, R_2, R_3, R_4 be four rewards such that $d^{\mathcal{R}}(R_1, R_2) < \epsilon$, $d^{\mathcal{R}}(R_3, R_4) < \epsilon$, and $d^{\mathcal{R}}(R_2, R_3) \gg \epsilon$ (or, alternatively, $R_1 \equiv_P R_2$, $R_3 \equiv_P R_4$, and $R_2 \not\equiv_P R_3$). Moreover, let $f, g : \mathcal{R} \to \Pi$ be two behavioural models where $f(R_1) = \pi_1$, $f(R_2) = f(R_3) = \pi_2$, $f(R_4) = \pi_3$, and $g(R_1) = g(R_2) = \pi_1$, $g(R_3) = g(R_4) = \pi_3$. We may assume that $f = g$ for all other reward functions. This is illustrated in the diagram below:



In this case, we have that $f(R_2) = f(R_3)$, but $d^{\mathcal{R}}(R_2, R_3) \gg \epsilon$. As such, f violates the third condition in Definition 8; a learning algorithm $\mathcal{L}$ based on f is *not* guaranteed to learn a reward function that has distance at most ϵ to the true reward function when there is no misspecification, because f cannot distinguish between R_2 and R_3, which have a large distance. However, if $f(R) = g(R')$, it does in this case follow that $d^{\mathcal{R}}(R, R') \leq \epsilon$. In other words, if the training data is coming from g, then a learning algorithm $\mathcal{L}$ based on f *is* guaranteed to learn a reward function that has distance at most ϵ to the true reward function. As such, we could define misspecification robustness in such a way that f would be considered to be robust to misspecification with g in this case. However, this seems unsatisfactory, because g essentially has to be carefully designed specifically to avoid certain blind spots in f. In other words, while condition 1 in Definition 7/8 is met, it is only met *spuriously*. The third condition is included to rule out these edge cases.

3.3 Intermediate Results About Our Definitions

In this section, we provide several lemmas and intermediate results about Definition 7 and 8. These results give insight into the properties of our problem setting, and will also be used to prove our object-level results. We begin by listing a number of interesting properties of Definition 7:

Lemma 11. *If f is not P-robust to misspecification with g, and* $\mathrm{Im}(g) \subseteq \mathrm{Im}(f)$*, then for any h, $h \circ f$ is not P-robust to misspecification with $h \circ g$.*

Proof. If f is not P-robust to misspecification with g, and $\mathrm{Im}(g) \subseteq \mathrm{Im}(f)$, then either $f \not\preceq P$, or $f = g$, or $f(R_1) = g(R_2)$ but $R_1 \not\equiv_P R_2$ for some R_1, R_2.

In the first case, if $f \not\preceq P$ then $h \circ f \not\preceq P$, as per Lemma 4. Thus $h \circ f$ is not P-robust to misspecification with any reward object (including $h \circ g$).

In the second case, if $f = g$ then $h \circ f = h \circ g$. This implies that $h \circ f$ is not P-robust to misspecification with $h \circ g$.

In the last case, suppose $f(R_1) = g(R_2)$ but $R_1 \not\equiv_P R_2$ for some R_1, R_2. Thus implies that $h \circ f(R_1) = h \circ g(R_2)$, but $R_1 \not\equiv_P R_2$. Thus $h \circ f$ is not P-robust to misspecification with $h \circ g$. □

Lemma 11 says that if f lacks robustness to a given form of misspecification, then any object that can be computed from f inherits a lack of robustness to its corresponding misspecification. This lemma can be seen as analogous to Lemma 4, and will later be used to show that broad classes of data models lack robustness to some forms of misspecification.

Lemma 12. *If f is P-robust to misspecification with g then* $\mathrm{Am}(g) \preceq P$.

Proof. Suppose f is P-robust to misspecification with g, and let R_1, R_2 be any two reward functions such that $g(R_1) = g(R_2)$. Since $\mathrm{Im}(g) \subseteq \mathrm{Im}(f)$ there is an R_3 such that $f(R_3) = g(R_1) = g(R_2)$. Since f is P-robust to misspecification with g, it must be the case that $R_3 \equiv_P R_1$ and $R_3 \equiv_P R_2$. By transitivity, we thus have that $R_1 \equiv_P R_2$. Since R_1 and R_2 were chosen arbitrarily, it must be that $R_1 \equiv_P R_2$ whenever $g(R_1) = g(R_2)$. □

It may be easier to understand Lemma 12 by considering the contrapositive statement; if $\mathrm{Am}(g) \not\preceq P$ then no f is P-robust to misspecification with g. In other words, if data from g is insufficient for identifying the P-class of the true reward function when there is no misspecification, then we cannot identify the correct P-class by using a misspecified data model. This means that we can never gain anything from misspecification.

Proposition 13. *If f is P-robust to misspecification with g and* $\mathrm{Im}(f) = \mathrm{Im}(g)$ *then g is P-robust to misspecification with f.*

Proof. If f is P-robust to misspecification with g then this immediately implies that $f \neq g$, and that if $f(R_1) = g(R_2)$ for some R_1, R_2 then $R_1 \equiv_P R_2$. Lemma 12 implies that $\mathrm{Am}(g) \preceq P$, and if $\mathrm{Im}(f) = \mathrm{Im}(g)$ then $\mathrm{Im}(f) \subseteq \mathrm{Im}(g)$. This means that g is P-robust to misspecification with f. □

Proposition 13 says that misspecification robustness is symmetric under many typical circumstances. For example, if f and g are both surjective, then $\mathrm{Im}(f) = \mathrm{Im}(g)$. This means that there are equivalence classes of behavioural models that are all robust to misspecification with each other.

Lemma 14. *Let* $\mathrm{Am}(f) \preceq P$. *Then there is no g such that f is P-robust to misspecification with g if and only if* $\mathrm{Am}(f) = P$.

Proof. First consider the case when $\mathrm{Am}(f) = P$, and assume for contradiction that f is P-robust to misspecification with g. Let R_1 be any reward function, and consider $g(R_1)$. Since $\mathrm{Im}(g) \subseteq \mathrm{Im}(f)$, there is an R_2 such that $f(R_2) = g(R_1)$. Since f is P-robust to misspecification with g, this implies that $R_2 \equiv_P R_1$. Moreover, if $\mathrm{Am}(f) = P$ then $R_2 \equiv_P R_1$ if and only if $f(R_2) = f(R_1)$, so it must be the case that $f(R_2) = f(R_1)$. Now, since $f(R_2) = f(R_1)$ and $f(R_2) = g(R_1)$, we have that $g(R_1) = f(R_1)$. Since R_1 was chosen arbitrarily, this implies that $f = g$, which is a contradiction. Hence, if $\mathrm{Am}(f) = P$ then there is no g such that f is P-robust to misspecification with g.

Next, consider the case when $\mathrm{Am}(f) \preceq P$ and $\mathrm{Am}(f) \neq P$. This implies that there are R_1, R_2 such that $R_1 \equiv_P R_2$ but $f(R_1) \neq f(R_2)$. We can then construct a g as follows; let $g(R_1) = f(R_2)$, $g(R_2) = f(R_1)$, and $g(R) = f(R)$ for all $R \notin \{R_1, R_2\}$. Now f is P-robust to misspecification with g. Hence, if $\mathrm{Am}(f) \preceq P$ and $\mathrm{Am}(f) \neq P$ then there is a g such that f is P-robust to misspecification with g, which in turn implies that if $\mathrm{Am}(f) \preceq P$ and there is no g such that f is P-robust to misspecification with g then $\mathrm{Am}(f) = P$. □

Lemma 14 has a few interesting implications. First of all, note that it means that we should expect most well-behaved data models to be robust to some forms of misspecification, assuming that $\text{Am}(f) \neq P$. Moreover, it also suggests that data models that are less ambiguous also are less robust to misspecification, and vice versa. One way to interpret this is to note that if $\text{Am}(f) \prec P$, then f is sensitive to properties of R that are irrelevant from the point of view of P. Specifically, it means that there are reward functions R_1, R_2 such that $f(R_1) \neq f(R_2)$ but $R_1 \equiv_P R_2$. Informally, we may then expect f to be robust to misspecification with g if f and g only differ in terms of such "irrelevant details" (c.f. Section 6.2).

Lemma 15. *Let* $\text{Am}(f) \preceq P$*. Then* f *is* P*-robust to misspecification with* g *if and only if* $f \neq g$ *and* $g = f \circ t$ *for some* $t : \mathcal{R} \to \mathcal{R}$ *such that* $R \equiv_P t(R)$ *for all* R*.*

Proof. First suppose that f is P-robust to misspecification with g — we will construct a t that fits our description. Since $\text{Im}(g) \subseteq \text{Im}(f)$, we have that there for each R exists an R' such that $g(R) = f(R')$. Let $t : \mathcal{R} \to \mathcal{R}$ be a function that maps each R to one such R'. Since by construction $g(R) = f(t(R))$ for each R, we have that $g = f \circ t$. Moreover, since f is P-robust to misspecification with g, we have that $R \equiv_P t(R)$. This completes the first direction.

For the other direction, suppose $f \neq g$ and $g = f \circ t$ for some $t : \mathcal{R} \to \mathcal{R}$ such that $R \equiv_P t(R)$ for all R. By assumption we have that $\text{Am}(f) \preceq P$. Moreover, we clearly have that $\text{Im}(g) \subseteq \text{Im}(f)$. Finally, if $g(R_1) = f(R_2)$ then $f(t(R_1)) = f(R_2)$. Since $\text{Am}(f) \preceq P$, this means that $t(R_1) \equiv_P R_2$. Moreover, since $R \equiv_P t(R)$ for all R, we have that $R_1 \equiv_P t(R_1)$. By transitivity, this means that $R_1 \equiv_P R_2$. Thus f is P-robust to misspecification with g, and we are done. □

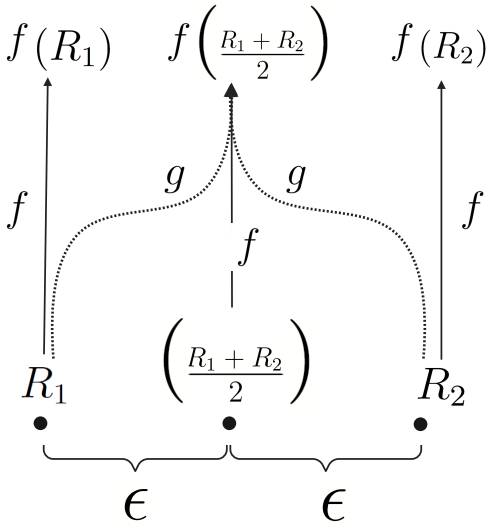
Lemma 15 is very important, because it provides us with an easy method for deriving necessary and sufficient conditions that completely describe what forms of misspecification any given data model f is robust to. In particular, given an equivalence relation P, if we can find the set T of all functions $t : \mathcal{R} \to \mathcal{R}$ such that $R \equiv_P t(R)$ for all R, then we can completely characterise the misspecification robustness of any data model f by simply composing f with each element of T.

We will later use this method to characterise the misspecification robustness of several important data models.

Let us next consider Definition 8. We will show that Definition 8 mostly fails to induce results analogous to those given in Lemma 11-15.

Proposition 16. *There exists a pseudometric $d^{\mathcal{R}}$ on $\mathcal{R}$ such that for each $\epsilon > 0$ there are reward objects $f, g : \mathcal{R} \to X$ where f is ϵ-robust to misspecification with g as measured by $d^{\mathcal{R}}$, but there are reward functions R_1, R_2 such that $g(R_1) = g(R_2)$ but $d^{\mathcal{R}}(R_1, R_2) > \epsilon$.*

Proof. For example, let $d^{\mathcal{R}}$ be the metric induced by the L_2-norm, let X be any set such that $|X| \geq |\mathcal{R}|$, and let $f : \mathcal{R} \to X$ be any injective function. Pick two reward functions R_1, R_2 such that $d^{\mathcal{R}}(R_1, R_2) = 2\epsilon$, let $g(R_1) = g(R_2) = f((R_1 + R_2)/2)$, and let $g(R) = f(R)$ for $R \neq R_1, R_2$.



□

As such, Definition 8 does not induce a result analogous to Lemma 12; there are f and g such that a learning algorithm that is based on f is guaranteed to learn a reward function that is close to the true reward function if trained on data generated from g, but where this is *not* true if we instead use a learning algorithm that is

based on g, even though the former algorithm is misspecified and the latter is not. This is somewhat pathological. However, we can prove a similar but weaker result:

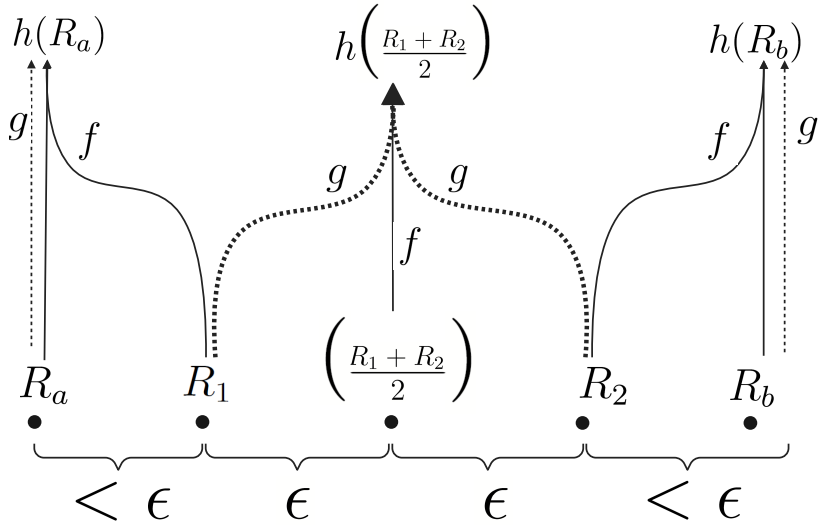
Lemma 17. *Let $f, g : \mathcal{R} \to$ be two reward objects, and let $d^{\mathcal{R}}$ be a pseudometric on $\mathcal{R}$. Suppose f is ϵ-robust to misspecification with g (as measured by $d^{\mathcal{R}}$). Then if $g(R_1) = g(R_2)$, we have that $d^{\mathcal{R}}(R_1, R_2) \leq 2\epsilon$.*

Proof. Let R_1, R_2 be any two reward functions such that $g(R_1) = g(R_2)$. From condition 2 in Definition 8, we have that there is a reward R_3 such that $f(R_3) = g(R_1) = g(R_2)$. From condition 1 in Definition 8, we have that $d^{\mathcal{R}}(R_3, R_1) \leq \epsilon$ and that $d^{\mathcal{R}}(R_3, R_2) \leq \epsilon$. The triangle inequality then implies that $d^{\mathcal{R}}(R_1, R_2) \leq 2\epsilon$. □

Definition 8 does also not induce a result analogous to Proposition 13:

Proposition 18. *There exists a pseudometric $d^{\mathcal{R}}$ on $\mathcal{R}$ such that for each $\epsilon > 0$ there are reward objects $f, g : \mathcal{R} \to X$ where f is ϵ-robust to misspecification with g as measured by $d^{\mathcal{R}}$, and $\mathrm{Im}(f) = \mathrm{Im}(g)$, but where g is not ϵ-robust to misspecification with f as measured by $d^{\mathcal{R}}$.*

Proof. For example, let $d^{\mathcal{R}}$ be the metric induced by the L_2-norm, let X be any set such that $|X| \geq |\mathcal{R}|$, and let $h : \mathcal{R} \to X$ be any injective function. Pick four reward functions R_1, R_2, R_a, R_b such that $d^{\mathcal{R}}(R_1, R_2) = 2\epsilon$, $d^{\mathcal{R}}(R_1, R_a) < \epsilon$, and $d^{\mathcal{R}}(R_2, R_b) < \epsilon$. Let $g(R_1) = g(R_2) = h((R_1 + R_2)/2)$, and let $g(R) = h(R)$ for $R \neq R_1, R_2$. Let $f(R_1) = h(R_a)$, $f(R_2) = h(R_b)$, and $f(R) = h(R)$ for $R \neq R_1, R_2$.

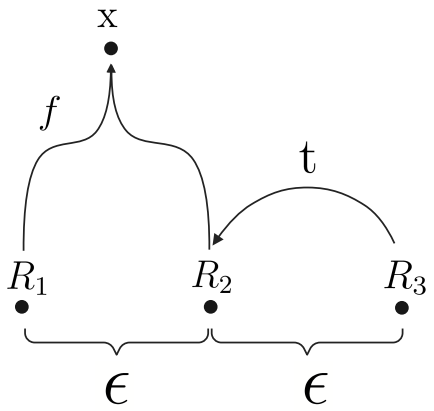


Now g is not ϵ-robust to misspecification with f, since $g(R_1) = g(R_2)$ even though $d^{\mathcal{R}}(R_1, R_2) = 2\epsilon$. However, f is ϵ-robust to misspecification with g. First, if $f(R) = g(R')$, then either $R = R'$, or $R' = (R_1 + R_2)/2$ and R is either R_1 or R_2. In the former case $d^{\mathcal{R}}(R, R') = 0$, and in the latter $d^{\mathcal{R}}(R_1, R_2) = \epsilon$. Moreover, if $f(R) = f(R')$, then either $R = R'$, or $R = R_1$ and $R' = R_a$ (or vice versa), or $R = R_2$ and $R' = R_b$ (or vice versa). In the first case $d^{\mathcal{R}}(R, R') = 0$, and in the latter two cases $d^{\mathcal{R}}(R, R') < \epsilon$. Next, $f \neq g$, since $f(R_1) \neq g(R_1)$ and $f(R_2) \neq g(R_2)$. Finally, $\mathrm{Im}(f) = \mathrm{Im}(g)$, since both $\mathrm{Im}(f)$ and $\mathrm{Im}(g)$ are equal to $\mathrm{Im}(h) \setminus \{h(R_1), h(R_2)\}$. □

In other words, Definition 8 fails to be symmetric even if $\mathrm{Im}(f) = \mathrm{Im}(g)$. This will make it more difficult to establish equivalence classes of behavioural models that are internally robust to misspecification. Lemma 14 cannot even be straightforwardly translated into Definition 8 for $\epsilon > 0$. Definition 8 also fails to induce a result analogous to Lemma 15:

Proposition 19. *There exists a pseudometric $d^{\mathcal{R}}$ on $\mathcal{R}$ such that for each $\epsilon > 0$ there is an $f : \mathcal{R} \to X$ such that if $f(R_1) = f(R_2)$ then $d^{\mathcal{R}}(R_1, R_2) \leq \epsilon$, and a $t : \mathcal{R} \to \mathcal{R}$ such that $d^{\mathcal{R}}(R, t(R)) \leq \epsilon$, where $f \neq f \circ t$, but where f is not ϵ-robust to misspecification with $f \circ t$ as measured by $d^{\mathcal{R}}$.*

Proof. For example, let $d^{\mathcal{R}}$ be the metric induced by the L_2-norm, and let X be any set such that $|X| \geq |\mathcal{R}|$. Pick three reward functions R_1, R_2, R_3 such that $d^{\mathcal{R}}(R_1, R_2) = \epsilon$, $d^{\mathcal{R}}(R_2, R_3) = \epsilon$, and $d^{\mathcal{R}}(R_1, R_3) = 2\epsilon$. Let f be injective, except that $f(R_1) = f(R_2)$, and let $t(R) = R$ for all R, except that $t(R_3) = R_2$. Now $f(R_1) = f \circ t(R_3)$, even though $d^{\mathcal{R}}(R_1, R_2) = 2\epsilon > \epsilon$, and so f is not ϵ-robust to misspecification with $f \circ t$ (as measured by $d^{\mathcal{R}}$).



□

This is particularly unfortunate, since Lemma 15 is very useful for easily deriving necessary and sufficient conditions for P-robustness. However, we can derive a *sufficient* condition for ϵ-robustness that mirrors Lemma 15; if f satisfies that $f(R_1) = f(R_2)$ implies $d^{\mathcal{R}}(R_1, R_2) \leq \epsilon$, and t satisfies that $d^{\mathcal{R}}(R, t(R)) \leq \epsilon$, then f is guaranteed to be 2ϵ-robust to misspecification with $f \circ t$. However, there can be g such that f is 2ϵ-robust to misspecification with g, but where g cannot be expressed in this way. We can also derive a necessary and sufficient condition by imposing stronger requirements on f:

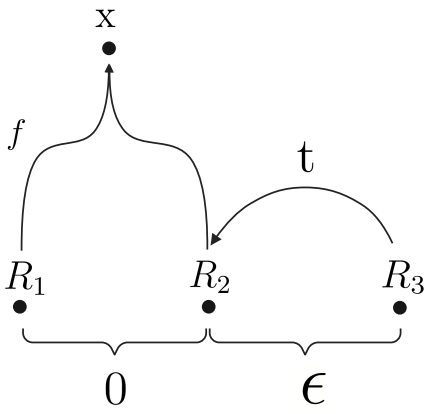
Lemma 20. *Let $f : \mathcal{R} \to X$ be a reward object, and let $d^{\mathcal{R}}$ be a pseudometric on $\mathcal{R}$. Assume that $f(R_1) = f(R_2) \implies d^{\mathcal{R}}(R_1, R_2) = 0$. Then f is ϵ-robust to*

misspecification with g as measured by $d^{\mathcal{R}}$ if and only if $g = f \circ t$ for some $t : \mathcal{R} \to \mathcal{R}$ such that $d^{\mathcal{R}}(R, t(R)) \leq \epsilon$ for all R, and such that $f \neq g$.

Proof. For the first direction, let $t : \mathcal{R} \to \mathcal{R}$ be a transformation such that $d^{\mathcal{R}}(R, t(R)) \leq \epsilon$ for all R, and let $g = f \circ t$. Suppose $f \neq g$. To show that f is ϵ-robust to misspecification with g, we need to show that:

1. If $f(R_1) = g(R_2)$ then $d^{\mathcal{R}}(R_1, R_2) \leq \epsilon$.
2. $\text{Im}(g) \subseteq \text{Im}(f)$.
3. If $f(R_1) = f(R_2)$ then $d^{\mathcal{R}}(R_1, R_2) \leq \epsilon$.
4. $f \neq g$.

For the first condition, suppose $f(R_1) = g(R_2)$, which implies that $f(R_1) = f \circ t(R_2)$. By assumption, we have that if $f(R) = f(R')$, then $d^{\mathcal{R}}(R, R') = 0$. This implies that $d^{\mathcal{R}}(R_1, t(R_2)) = 0$. Moreover, we have that $d^{\mathcal{R}}(R, t(R)) \leq \epsilon$ for all R; this implies that $d^{\mathcal{R}}(R_2, t(R_2)) \leq \epsilon$. By the triangle inequality, we then have that $d^{\mathcal{R}}(R_1, R_2) \leq 0 + \epsilon = \epsilon$. Since R_1 and R_2 were chosen arbitrarily, this means that condition 1 holds. Condition 2 holds straightforwardly, from the construction of g. For condition 3, note that we by assumption have that if $f(R_1) = f(R_2)$, then $d^{\mathcal{R}}(R_1, R_2) = 0 < \epsilon$. Condition 4 is satisfied by direct assumption. This proves the first direction.



For the other direction, let f be ϵ-robust to misspecification with g (as measured by $d^{\mathcal{R}}$). Since $\text{Im}(g) \subseteq \text{Im}(f)$, we have that there for each R exists an R' such that $g(R) = f(R')$. Let $t : \mathcal{R} \to \mathcal{R}$ be a function that maps each R to one such R'. Since by construction $g(R) = f(t(R))$ for each R, we have that $g = f \circ t$. Moreover, since f is ϵ-robust to misspecification with g as measured by $d^{\mathcal{R}}$, we have that $d^{\mathcal{R}}(R, t(R)) \leq \epsilon$. This completes the proof of the other direction. □

Moreover, Definition 8 *does* induce a result analogous to Lemma 11:

Proposition 21. *For any pseudometric $d^{\mathcal{R}}$ on $\mathcal{R}$ and any $\epsilon \geq 0$, if f is not ϵ-robust to misspecification with g as measured by $d^{\mathcal{R}}$, and $\text{Im}(g) \subseteq \text{Im}(f)$, then for any h, $h \circ f$ is not ϵ-robust to misspecification with $h \circ g$ as measured by $d^{\mathcal{R}}$.*

Proof. If f is not ϵ-robust to misspecification with g as measured by $d^{\mathcal{R}}$, and $\text{Im}(g) \subseteq \text{Im}(f)$, then either there are R_1, R_2 such that $f(R_1) = g(R_2)$ but $d^{\mathcal{R}}(R_1, R_2) > \epsilon$, or there are R_1, R_2 such that $f(R_1) = f(R_2)$ but $d^{\mathcal{R}}(R_1, R_2) > \epsilon$, or $f = g$.

In the first case, if $f(R_1) = g(R_2)$ but $d^{\mathcal{R}}(R_1, R_2) > \epsilon$ then $h \circ f(R_1) = h \circ g(R_2)$ but $d^{\mathcal{R}}(R_1, R_2) > \epsilon$. In the second case, if $f(R_1) = f(R_2)$ but $d^{\mathcal{R}}(R_1, R_2) > \epsilon$ then $h \circ f(R_1) = h \circ f(R_2)$ but $d^{\mathcal{R}}(R_1, R_2) > \epsilon$. In the third case, if $f = g$ then $h \circ f = h \circ g$. In each case, we thus have that $h \circ f$ is not ϵ-robust to misspecification with $h \circ g$ as measured by $d^{\mathcal{R}}$. □

The comparison between Lemma 11-15 and the results provided above exemplify the fact that Definition 7 sometimes lets us derive results that are stronger and more informative than what is possible using Definition 8, unless we assume that $\epsilon = 0$ (in which case Definitions 7 and 8 are equivalent). For this reason, we will make use of both Definition 7 and Definition 8.

3.4 Reward Transformations

Recall that a *reward transformation* is a map $t : \mathcal{R} \to \mathcal{R}$. In this section, we introduce several important classes of reward transformations, that we will later use to express our results. First recall *potential shaping*, which was first introduced by Ng, Harada, and Russell, 1999:

Definition 22. A *potential function* is a function $\Phi : \mathcal{S} \to \mathbb{R}$. Given a discount γ, we say that R_1 and R_2 differ by *potential shaping* with γ if for some potential Φ,

$$R_2(s, a, s') = R_1(s, a, s') + \gamma \cdot \Phi(s') - \Phi(s).$$

for all $s, s' \in \mathcal{S}$ and all $a \in \mathcal{A}$.

Ng, Harada, and Russell, 1999 proved that if R_1 and R_2 differ by potential shaping, then they have the same optimal policies for any choice of τ and μ_0. Potential shaping also has many other interesting properties, which we discuss in more detail in Section 4.1. We next define two new classes of transformations, starting with *S'-redistribution*.

Definition 23. Given a transition function τ, we say that R_1 and R_2 differ by *S'-redistribution* with τ if

$$\mathbb{E}_{S' \sim \tau(s,a)} [R_1(s, a, S')] = \mathbb{E}_{S' \sim \tau(s,a)} [R_2(s, a, S')]$$

for all $s, s' \in \mathcal{S}$ and all $a \in \mathcal{A}$.

S'-redistribution accounts for any difference between R_1 and R_2 that does not affect the expected reward. If $s_1,\ s_2 \in \text{Supp}(\tau(s, a))$ then S'-redistribution can increase $R(s, a, s_1)$ if it decreases $R(s, a, s_2)$ proportionally. S'-redistribution can also change R arbitrarily for transitions that occur with probability 0. We next consider *optimality-preserving transformations*:

Definition 24. Given a transition function τ and a discount γ, we say that R_1 and R_2 differ by an *optimality-preserving transformation* with τ and γ if there exists a function $\psi : \mathcal{S} \to \mathbb{R}$ such that

$$\mathbb{E}_{S' \sim \tau(s,a)}[R_2(s, a, S') + \gamma \cdot \psi(S')] \leq \psi(s),$$

with equality if and only if $a \in \text{argmax}_{a \in \mathcal{A}} A^\star_1(s, a)$.

As the name suggests, an optimality-preserving transformation preserves optimal policies (c.f. Theorem 34). Intuitively speaking, an optimality-preserving transformation let us pick an arbitrary new value function ψ, and then adjust R_2 in any way that respects the new value function and the argmax of $A_1^\star$ — the latter condition ensures that the same actions (and hence the same policies) stay optimal.

In addition to these transformations, we also say that R_1 and R_2 differ by *positive linear scaling* if $R_2 = c \cdot R_1$ for some positive constant c, and that they differ by *constant shift* if $R_2 = R_1 + c$ for some constant c. Based on these definitions, we can now specify several *sets* of reward transformations:

1. Let PS_γ be the set of all reward transformations t such that $t(R)$ is given by potential shaping of R relative to the discount γ.

2. Let $S'\mathrm{R}_\tau$ be the set of all reward transformations t such that $t(R)$ is given by S'-redistribution of R relative to the transition function τ.

3. Let LS be the set of all reward transformations t that scale each reward function by some positive constant, i.e. for each R there is a $c \in \mathbb{R}^+$ such that $t(R)(s, a, s') = c \cdot R(s, a, s')$.

4. Let CS be the set of all reward transformations t that shift each reward function by some constant, i.e. for each R there is a $c \in \mathbb{R}$ such that $t(R)(s, a, s') = R(s, a, s') + c$.

5. Let $\mathrm{OP}_{\tau,\gamma}$ be the set of all reward transformations t such that $t(R)$ is given by an optimality-preserving transformation of R with τ and γ.

Note that these sets are defined in a way that allows their transformations to be "sensitive" to the reward function it takes as input. For example, a transformation $t \in \mathrm{PS}_\gamma$ might apply one potential function Φ_1 to R_1, and a different potential function Φ_2 to R_2, and a transformation $t \in \mathrm{LS}$ might scale R_1 by a positive constant c_1, and R_2 by a different constant c_2, etc. Many of our results will be expressed in terms of these sets.

It is worth noting that CS $\subseteq$ PS$_\gamma$; to see this, note that we for any constant c and any discount factor γ can define a potential function Φ such that $\Phi(s) = c/(\gamma - 1)$ for all states s. Moreover, each of PS$_\gamma$, S'R$_\tau$, LS and CS are subsets of OP$_{\tau,\gamma}$. First note that OP$_{\tau,\gamma}$ is exactly the set of all reward transformations that preserve optimal policies (c.f. Theorem 34). Next, it should be clear that positive linear scaling of the reward preserves the set of optimal policies, which means that LS $\subseteq$ OP$_{\tau,\gamma}$. Moreover, using the linearity of expectation, it is also easy to see that S'-redistribution preserves optimal policies, which means that S'R$_\tau \subseteq$ OP$_{\tau,\gamma}$. Finally, Ng, Harada, and Russell, 1999 show that potential shaping preserves optimal policies, which means that CS $\subseteq$ PS$_\gamma \subseteq$ OP$_{\tau,\gamma}$ (c.f. also Proposition 29).

We will also combine sets of reward transformations to form bigger sets. Specifically, if T_1 and T_2 are sets of reward transformations, then we use $T_1 \bigodot T_2$ to denote the set of all transformations that can be obtained by composing transformations in T_1 and T_2 arbitrarily, in any order. Formally, we define this operator in the following way:

Definition 25. Let T_1 and T_2 be two (non-empty) sets of reward transformations. Define S_0 as $T_1 \cup T_2$, and

$$S_{i+1} = \{t_1 \circ t_2 : t_1 \in T_1 \cup T_2, t_2 \in S_i\}.$$

Then $T_1 \bigodot T_2 = \bigcup_{i=0}^{\infty}\{S_i : i \in \mathbb{N}\}$.

For example, this means that PS$_\gamma \bigodot S'$R$_\tau$ is the set of all reward transformations that can be created by composing potential shaping and S'-redistribution, in any order. In the text, we will sometimes refer to this set as "potential shaping and S'-redistribution". This means that the statement "R_1 and R_2 differ by potential shaping and S'-redistribution" should be understood as saying that there is a $t \in$ PS$_\gamma \bigodot S'$R$_\tau$ such that $R_2 = t(R_1)$, and so on. Also note that $\bigodot$ is both commutative and associative.

We next note a few basic algebraic properties of our transformations:

Proposition 26. *If T is* PS$_\gamma$, S'R$_\tau$, LS, CS, *or* OP$_{\tau,\gamma}$, *then*

1. *The identity transformation,* id*, is in* T*.*

2. *For all* $t \in T$ *there is a* $t^- \in T$ *such that* $t \circ t^- = \mathrm{id}$*.*

3. *For all* $t, t' \in T$*, we have that* $t \circ t' \in T$*.*

Proof. For (1), first note that id satisfies the conditions for potential shaping with the function Φ such that $\Phi(s) = 0$ for all s; hence $\mathrm{id} \in \mathrm{PS}_\gamma$. Next, since trivially $\mathbb{E}_{S' \sim \tau(s,a)}[R(s,a,S')] = \mathbb{E}_{S' \sim \tau(s,a)}[R(s,a,S')]$, we have that $\mathrm{id} \in S'\mathrm{R}_\tau$. Moreover, id satisfies the conditions for positive linear scaling with a factor $c = 1$, and so $\mathrm{id} \in \mathrm{LS}$. Furthermore, id satisfies the conditions for constant shift with a factor $c = 0$, and so $\mathrm{id} \in \mathrm{CS}$. Finally, id satisfies the conditions for optimality-preserving transformations, where $\Psi = V^\star$ (this is precisely the Bellman optimality equation for $V^\star$, see Equation 2.3).

For (2), first note that if R_1 and R_2 differ by potential shaping with Φ, then R_2 and R_1 differ by potential shaping with $-\Phi$. Furthermore, we trivially have that if $\mathbb{E}_{S' \sim \tau(s,a)}[R_1(s,a,S')] = \mathbb{E}_{S' \sim \tau(s,a)}[R_2(s,a,S')]$ then $\mathbb{E}_{S' \sim \tau(s,a)}[R_2(s,a,S')] = \mathbb{E}_{S' \sim \tau(s,a)}[R_1(s,a,S')]$. Moreover, if R_1 and R_2 differ by positive linear scaling by c, then R_2 and R_1 differ by positive linear scaling by $(1/c)$. Similarly, if R_1 and R_2 differ by constant shift with c, then R_2 and R_1 differ by constant shift with $-c$. Finally, if R_1 and R_2 differ by an optimality-preserving transformation, then $A_1^\star = A_2^\star$, and so R_2 and R_1 also differ by an optimality-preserving transformation.

For (3), note that if R_2 is given by potential shaping of R_1 with Φ_1, and R_3 is given by potential shaping of R_2 with Φ_2, then R_3 is given by potential shaping of R_1 with $\Phi_1 + \Phi_2$. Moreover, we trivially have that

$$\mathbb{E}_{S' \sim \tau(s,a)}[R_1(s,a,S')] = \mathbb{E}_{S' \sim \tau(s,a)}[R_3(s,a,S')]$$

if

$$\mathbb{E}_{S' \sim \tau(s,a)}[R_1(s,a,S')] = \mathbb{E}_{S' \sim \tau(s,a)}[R_2(s,a,S')]$$

and

$$\mathbb{E}_{S' \sim \tau(s,a)}[R_2(s,a,S')] = \mathbb{E}_{S' \sim \tau(s,a)}[R_3(s,a,S')]\,.$$

Next, if R_2 is given by positive linear scaling of R_1 with c_1, and R_3 is given by positive linear scaling of R_2 with c_2, then R_3 is given by positive linear scaling of R_1 with $c_1 \cdot c_2$. Similarly, if R_2 is given by constant shift of R_1 with c_1, and R_3 is given by constant shift of R_2 with c_2, then R_3 is given by constant shift of R_1 with $c_1 + c_2$. Finally, suppose that R_1 and R_2 differ by an optimality-preserving transformation, and that R_2 and R_3 differ by an optimality-preserving transformation. We then have that $A_1^\star = A_2^\star$ and $A_2^\star = A_3^\star$, so $A_1^\star = A_3^\star$, which means that R_1 and R_3 differ by an optimality-preserving transformation. □

Proposition 26 implies that each of the sets PS_γ, $S'\mathrm{R}_\tau$, LS, CS, and $\mathrm{OP}_{\tau,\gamma}$ form groups. It also implies that each of these sets partitions $\mathcal{R}$ into equivalence classes. Note that these properties do not hold for arbitrary sets of reward transformations, so they are special properties of PS_γ, $S'\mathrm{R}_\tau$, LS, CS, and $\mathrm{OP}_{\tau,\gamma}$. The following is also worth noting:

Proposition 27. *Let T_1 and T_2 be sets of reward transformations such that if T is T_1 or T_2, then*

1. *The identity transformation,* id*, is in T.*
2. *For all $t \in T$ there is a $t^- \in T$ such that $t \circ t^- = \mathrm{id}$.*

We then have that

1. *The identity transformation,* id*, is in $T_1 \bigodot T_2$.*
2. *For all $t \in T_1 \bigodot T_2$ there is a $t^- \in T_1 \bigodot T_2$ such that $t \circ t^- = \mathrm{id}$.*
3. *For all $t, t' \in T_1 \bigodot T_2$, we have that $t \circ t' \in T_1 \bigodot T_2$.*

Proof. For (1), note that $T_1, T_2 \subset T_1 \bigodot T_2$. For (2), note that if $t \in T_1 \bigodot T_2$, then $t = t_1 \circ \cdots \circ t_n$, where each transformation t_i is in either T_1 or T_2. Let $t^- = t_n^- \circ \cdots \circ t_1^-$. Now $t \circ t^- = \mathrm{id}$, and $t^- \in T_1 \bigodot T_2$. It is immediate from the definition of the $\bigodot$-operator that (3) is satisfied. □

This means that the properties described in Proposition 26 also hold for any set of reward transformations which can be constructed from PS_γ, $S'\mathrm{R}_\tau$, LS, CS, and $\mathrm{OP}_{\tau,\gamma}$ using the $\odot$-operator. The following is also useful:

Proposition 28. *If both T_1 and T_2 are PS_γ, $S'\mathrm{R}_\tau$, LS, CS, or $\mathrm{OP}_{\tau,\gamma}$, then for each $t_1 \in T_1$ and $t_2 \in T_2$, there is a $t_1' \in T_1$ and a $t_2' \in T_2$ such that $t_1 \circ t_2 = t_2' \circ t_1'$.*

Proof. If $T_1 = T_2$, the proposition is trivial. Next, recall that each of PS_γ, $S'\mathrm{R}_\tau$, LS, and CS is a subset of $\mathrm{OP}_{\tau,\gamma}$. This means that if one of T_1 or T_2 is $\mathrm{OP}_{\tau,\gamma}$, then we can set $t_2' = t_1 \circ t_2$ and $t_1' = \mathrm{id}$, or vice versa (recalling also the properties listed in Proposition 26).

For the remaining cases, let S be the set of all reward functions that can be expressed as $R(s,a,s') = \gamma \cdot \Phi(s') - \Phi(s)$ for some potential function Φ, and let Z be the set of all reward functions that satisfy $\mathbb{E}_{S' \sim \tau(s,a)}\left[R(s,a,S')\right] = 0$. Note that R_1 and R_2 differ by potential shaping if and only if $R_1 = R_2 + R_S$ for some $R_S \in S$, and that R_1 and R_2 differ by S'-redistribution if and only if $R_1 = R_2 + R_Z$ for some $R_Z \in Z$.

Now, let T_1 be PS_γ and T_2 be $S'\mathrm{R}_\tau$. We now have that for all R, there is an $R_S \in S$ and an $R_Z \in Z$ such that $t_1 \circ t_2(R) = R + R_S + R_Z$. Since vector addition is commutative, this means that we can find the desired t_1' and t_2'. The case where T_1 is $S'\mathrm{R}_\tau$ and T_2 is PS_γ is analogous.

Next, let $T_1 = \mathrm{PS}_\gamma$ and $T_2 = \mathrm{LS}$. We now have that there for all R is an $R_S \in S$ and a $c \in \mathbb{R}^+$ such that $t_1 \circ t_2(R) = c \cdot R + R_S$. This also means that $t_1 \circ t_2(R) = c \cdot (R + \frac{1}{c} R_S)$. Since $\frac{1}{c} R_S \in S$, this means that we can find the desired t_1' and t_2'. The case where $T_1 = \mathrm{LS}$ and $T_2 = \mathrm{PS}_\gamma$ is analogous, and likewise for the case where T_1 and T_2 are $S'\mathrm{R}_\tau$ and LS.

The case where T_1 or T_2 is CS is covered by the above cases, since $\mathrm{CS} \subseteq \mathrm{PS}_\gamma$. This completes the proof. □

Proposition 28 means that we do not have to be very careful about the order in which transformations from PS_γ, $S'\mathrm{R}_\tau$, LS, CS, or $\mathrm{OP}_{\tau,\gamma}$ are applied. For example, if we can produce R_1 from R_2 by first applying potential shaping, and then applying

S'-redistribution, then we can also do this by first applying S'-redistribution, and then applying potential shaping, and so on. This fact, combined with the properties listed in Proposition 26, will substantially reduce the number of cases that we have to consider in some proofs. For example, if $t \in \mathrm{PS}_\gamma \bigodot \mathrm{LS} \bigodot S'\mathrm{R}_\tau$, then it can always be expressed as $t_1 \circ t_2 \circ t_3$, where $t_1 \in \mathrm{PS}_\gamma$, $t_2 \in \mathrm{LS}$, and $t_3 \in S'\mathrm{R}_\tau$, etc. We will make use of these properties in many of our proofs, sometimes without explicitly stating that we are doing so.

In Section 4.1, we list and prove several key properties of the reward transformations we have introduced above. These properties are primarily used in our proofs, but may also be helpful for gaining an intuitive understanding of how these reward transformations work.

3.5 Behavioural Models

In this section, we will introduce special notation for the three behavioural models that are most common in the current IRL literature, i.e. optimal policies, Boltzmann-rational policies, and MCE policies. Given a transition function τ and a discount parameter γ, let $b_{\tau,\gamma,\beta} : \mathcal{R} \to \Pi$ be the function that returns the Boltzmann-rational policy of R with temperature β, and let $c_{\tau,\gamma,\alpha} : \mathcal{R} \to \Pi$ be the function that returns the MCE policy of R with weight α. These policies exist and are unique for each τ, γ, β, and α, and so $b_{\tau,\gamma,\beta}$ and $c_{\tau,\gamma,\alpha}$ are well-defined.

The optimality model requires a bit more care, because there may in general be more than one policy that is optimal under a given reward function. To resolve this, recall that a policy is optimal if and only if it only gives support to optimal actions, where the "optimal actions" are the actions that maximise $Q^\star$. A state may have multiple optimal actions, so we can get multiple optimal policies by breaking ties in different ways. However, if an optimal policy gives support to multiple actions in some state, then we would normally not expect the exact probability it assigns to each action to convey any information about the reward function. We will therefore only look at the actions that the optimal policy takes, and ignore the relative probability it assigns to those actions. Formally, we will treat optimal

policies as functions $\pi_\star : \mathcal{S} \to \mathcal{P}(\text{argmax}_{a \in \mathcal{A}} A^\star) - \{\varnothing\}$; i.e. as functions that for each state return a non-empty subset of the set of all actions that are optimal in that state. Let $\mathcal{O}_{\tau,\gamma}$ be the set of all functions that return such policies (relative to transition function τ and discount factor γ). Moreover, let $o^\star_{\tau,\gamma} \in \mathcal{O}_{\tau,\gamma}$ be the function that, given R, returns the function that maps each state to the set of *all* actions which are optimal in that state. Intuitively, $o^\star_{\tau,\gamma}$ corresponds to optimal policies that take all optimal actions with positive probability. Alternatively, we can also think of $o^\star_{\tau,\gamma}$ as corresponding to the set of all optimal policies (noting that this set determines the set of optimal actions for each state, and vice versa).

Any observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes.

— Charles Goodhart, 1975.

4 Comparing Reward Functions

When analysing a reward learning algorithm, we wish to derive claims that compare the learnt reward function to the underlying true reward function, given different setups and conditions. To do this, we must first have principled methods for comparing reward functions. In this chapter, we discuss different methods for doing this. First, we will introduce two natural *equivalence relations* on the space of all reward functions, and characterise the corresponding equivalence classes. We will also introduce a family of *pseudometrics* on the space of all reward functions, and show that these pseudometrics satisfy several desirable properties. In later chapters, we will use these reward transformations, equivalence classes, and metrics, to express and prove our results about IRL algorithns.

4.1 Key Properties of Reward Transformations

In this section, we will list and prove a few key properties of the reward transformations we introduced in Section 3.4. These properties will help to provide some intuition for how these transformations work, and will also be used to prove some of our later results.

We first prove a number of important properties of potential shaping, which are not explicitly discussed in Ng, Harada, and Russell, 1999. These properties will be

important for our later results, and will also help with providing more intuition for what potential shaping does and how it behaves.

Proposition 29. *Let R_1 and R_2 be any two reward functions. If R_2 is produced by potential shaping of R_1 with a potential function Φ, then*

1. $G_2(\xi) = G_1(\xi) - \Phi(s_0)$,
2. $Q_2^\pi(s,a) = Q_1^\pi(s,a) - \Phi(s)$,
3. $V_2^\pi(s) = V_1^\pi(s) - \Phi(s)$,
4. $A_2^\pi(s,a) = A_1^\pi(s,a)$, *and*
5. $J_2(\pi) = J_1(\pi) - \mathbb{E}_{S_0 \sim \mu_0}[\Phi(S_0)]$

for all trajectories ξ, policies π, states s, actions a, transition functions τ, and initial state distributions μ_0. In (1), s_0 is the first state of ξ.

Proof. To prove (1), first consider a finite trajectory fragment ζ with n transitions. It is then easy to prove via induction on n that $G_2(\zeta) = G_1(\zeta) + \gamma^n \cdot \Phi(s_n) - \Phi(s_0)$, where s_0 is the first state of ζ, and s_n is the last state. Moreover, Φ is bounded (since $\mathcal{S}$ is finite) and $\gamma \in (0,1)$. This means that $\gamma^n \cdot \Phi(s_n)$ goes to 0 as n goes to infinity. (2) and (3) follow immediately from (1). For (4), note that $A^\pi(s,a) = Q^\pi(s,a) - V^\pi(s)$. (5) is immediate from (3). □

This means that we can think of Φ as assigning "credit" to each state s, such that the total reward of any policy or trajectory which starts in that state s will lose a total of $\Phi(s)$ reward. Note that this directly implies that potential shaping preserves optimal policies (and the ordering of policies), since the value of every policy is shifted by the same amount. Moreover, this property also extends to the soft Q-function:

Proposition 30. *Let R_1 and R_2 be two reward functions, where R_2 is given by potential shaping of R_1 with Φ. Let $Q_{\alpha,1}^{\mathrm{S}}$ and $Q_{\alpha,2}^{\mathrm{S}}$ be their soft Q-functions (for some τ, γ, and α). Then $Q_{\alpha,2}^{\mathrm{S}}(s,a) = Q_{\alpha,1}^{\mathrm{S}}(s,a) - \Phi(s)$.*

Proof. Recall that $Q^{\mathrm{S}}_{\alpha,1}$ is the unique function which satisfies

$$Q^{\mathrm{S}}_{\alpha,1}(s,a) = \mathbb{E}_{S' \sim \tau(s,a)} \left[R_1(s,a,S') + \gamma\alpha \log \sum_{a' \in \mathcal{A}} \exp\left(\left(\frac{1}{\alpha}\right) Q^{\mathrm{S}}_{\alpha,1}(S',a') \right) \right]$$

for all s, a. Since $R_2(s,a,s') = R_1(s,a,s') + \gamma \cdot \Phi(s') - \Phi(s)$, we can rewrite the right-hand side of this equation as

$$\begin{aligned} &\mathbb{E}\left[R_1(s,a,S') + \gamma\alpha \log \sum_{a' \in \mathcal{A}} \exp\left(\left(\frac{1}{\alpha}\right) Q^{\mathrm{S}}_{\alpha,1}(S',a') \right) \right] \\ =&\mathbb{E}\left[R_2(s,a,S') - \gamma \cdot \Phi(S') + \Phi(s) + \gamma\alpha \log \sum_{a' \in \mathcal{A}} \exp\left(\left(\frac{1}{\alpha}\right) Q^{\mathrm{S}}_{\alpha,1}(S',a') \right) \right] \end{aligned}$$

By moving $\Phi(s)$ to the left-hand side, we get that $Q^{\mathrm{S}}_{\alpha,1}(s,a) - \Phi(s)$ is equal to

$$\begin{aligned} &\mathbb{E}\left[R_2(s,a,S') - \gamma \cdot \Phi(S') + \gamma\alpha \log \sum_{a' \in \mathcal{A}} \exp\left(\left(\frac{1}{\alpha}\right) Q^{\mathrm{S}}_{\alpha,1}(S',a') \right) \right] \\ =&\mathbb{E}\left[R_2(s,a,S') + \gamma\alpha \left(-\left(\frac{1}{\alpha}\right) \Phi(S') + \log \sum_{a' \in \mathcal{A}} \exp\left(\left(\frac{1}{\alpha}\right) Q^{\mathrm{S}}_{\alpha,1}(S',a') \right) \right) \right] \\ =&\mathbb{E}\left[R_2(s,a,S') + \gamma\alpha \left(\log\exp -\left(\frac{1}{\alpha}\right) \Phi(S') + \log \sum_{a' \in \mathcal{A}} \exp\left(\left(\frac{1}{\alpha}\right) Q^{\mathrm{S}}_{\alpha,1}(S',a') \right) \right) \right] \\ =&\mathbb{E}\left[R_2(s,a,S') + \gamma\alpha \log \left(\left(\exp -\left(\frac{1}{\alpha}\right) \Phi(S') \right) \left(\sum_{a' \in \mathcal{A}} \exp\left(\left(\frac{1}{\alpha}\right) Q^{\mathrm{S}}_{\alpha,1}(S',a') \right) \right) \right) \right] \\ =&\mathbb{E}\left[R_2(s,a,S') + \gamma\alpha \log \sum_{a' \in \mathcal{A}} \exp\left(\left(\frac{1}{\alpha}\right) (Q^{\mathrm{S}}_{\alpha,1}(S',a') - \Phi(S')) \right) \right] \end{aligned}$$

This means that $Q^{\mathrm{S}}_{\alpha,1}(s,a) - \Phi(s)$ satisfies the soft Q-function recursion (Equation 2.5) for R_2. Since the soft Q-function is the unique solution to this equation, we conclude that $Q^{\mathrm{S}}_{\alpha,2}(s,a) = Q^{\mathrm{S}}_{\alpha,1}(s,a) - \Phi(s)$. □

We next show that potential shaping of the reward function R correspond *exactly* to constant shift of the return function, G:

Proposition 31. *Let R_1 be any reward function and k any constant. Then we have that $G_2(\xi) = G_1(\xi) + k$ for all trajectories that start in a state s if and only if R_2 is given by potential shaping of R_1 with a potential function Φ such that $\Phi(s) = -k$.*

Proof. The first direction follows from part (1) of Proposition 29. For the other direction, suppose $G_2(\xi) = G_1(\xi) + k$ for all trajectories that start in state s. We will show that this implies a constant difference between G_1 and G_2 for all trajectories starting in *any* state, and then use this difference to define a potential function that transforms R_1 into R_2.

Consider an arbitrary state $s' \in \mathcal{S}$. Given a trajectory ξ starting in s', let $\Delta_\xi = G_2(\xi) - G_1(\xi)$. We will show that for any two trajectories ξ_1, ξ_2 starting in s', we have that $\Delta_{\xi_1} = \Delta_{\xi_2}$. Let ζ be a finite trajectory fragment that starts in s and ends in s', and let $n = |\zeta|$. Let $\zeta + \xi$ denote the concatenation of ζ and ξ. Then,

$$\begin{aligned}\Delta_\xi &= G_2(\xi) - G_1(\xi) \\ &= \frac{G_2(\zeta + \xi) - G_2(\zeta)}{\gamma^n} - \frac{G_1(\zeta + \xi) - G_1(\zeta)}{\gamma^n} \\ &= \frac{k - G_2(\zeta) + G_1(\zeta)}{\gamma^n}\,.\end{aligned}$$

The first line follows from the fact that $G(\zeta + \xi) = G(\zeta) + \gamma^{|\zeta|} G(\xi)$. For the second line, note that $\zeta + \xi$ is a trajectory starting in s. Thus, by assumption, we have that $G_2(\zeta + \xi) - G_1(\zeta + \xi) = k$. Since this expression is independent of ξ, this means that $\Delta_{\xi_1} = \Delta_{\xi_2}$ for any two trajectories ξ_1, ξ_2 starting in s'. Since s' was picked arbitrarily, this holds for all states s'.

Let $\Phi : \mathcal{S} \to \mathbb{R}$ be the potential function where $\Phi(s')$ is the value of $-\Delta_\xi$ for all trajectories ξ which start in s'. In other words, $\Phi(s') = G_1(\xi) - G_2(\xi)$ for all trajectories ξ which start in s'. We will show that R_2 is given by potential shaping of R_1 with Φ. Let (s, a, s') be any transition, let ξ' be any trajectory starting in s', and let $\xi = (s, a, s') + \xi'$. Then:

$$\begin{aligned}&R_1(s, a, s') + \gamma\Phi(s') - \Phi(s) \\ =&R_1(s, a, s') + \gamma(G_1(\xi') - G_2(\xi')) - (G_1(\xi) - G_2(\xi)) \\ =&G_2(\xi) - \gamma G_2(\xi') + R_1(s, a, s') + \gamma G_1(\xi') - G_1(\xi) \\ =&G_2(\xi) - \gamma G_2(\xi') + G_1(\xi) - G_1(\xi) \\ =&R_2(s, a, s')\,.\end{aligned}$$

Thus, R_2 is given by potential shaping of R_1 with Φ. Finally, note that $\Phi(s) = -k$, since $\Phi(s) = G_1(\xi) - G_2(\xi)$ for all trajectories starting in s, and since $G_2(\xi) = G_1(\xi) + k$ for all trajectories that start in s. This completes the proof. □

Note that Proposition 31 quantifies over all trajectories in $(\mathcal{S}\times\mathcal{A})^\omega$, rather than all trajectories which are possible under some transition function τ. However, it should be clear from Proposition 31 that $G_2(\xi) = G_1(\xi) + k$ for all *possible* trajectories that start in a state s if and only if R_2 is given by potential shaping of R_1 with a potential function Φ such that $\Phi(s) = -k$, and an arbitrary change of all transitions that are unreachable from s. Next, we show that positive linear scaling of G corresponds to a combination of potential shaping and positive linear scaling of R:

Proposition 32. *$G_2(\xi) = c \cdot G_1(\xi)$ for all trajectories ξ that start in a state s if and only if R_2 is given by potential shaping of R_1 with a potential function Φ such that $\Phi(s) = 0$, and positive linear scaling by a factor of c.*

Proof. For the first direction, suppose R_2 is given by potential shaping of R_1 with a potential function Φ such that $\Phi(s) = 0$, and positive linear scaling by a factor of c. Let R_3 be given by potential shaping of R_1 with Φ, so that $R_2 = c \cdot R_3$. As per Proposition 29, this means that $G_3(\xi) = G_1(\xi)$ for all trajectories ξ that start in s. This then implies that $G_2(\xi) = c \cdot G_1(\xi)$ for all trajectories ξ that start in a state s. This completes the first direction.

For the other direction, suppose $G_2(\xi) = c \cdot G_1(\xi)$ for all trajectories ξ that start in a state s. Let R_c be the reward function given by $c \cdot R_1$. It is clear that $G_c(\xi) = c \cdot G_1(\xi)$. Thus, $G_2(\xi) = c \cdot G_1(\xi)$ for all trajectories ξ that start in a state s, if and only if $G_2(\xi) = G_c(\xi)$ for all trajectories ξ that start in a state s. As per Proposition 31, this is equivalent to R_2 being produced from R_c by potential shaping with a potential function Φ such that $\Phi(s) = 0$. This means that we can produce R_2 from R_1 by first applying positive linear scaling by a factor of c, and then applying potential shaping. This completes the other direction. □

Together, Proposition 31 and 32 imply that potential shaping and positive linear scaling of R correspond exactly to affine transformations of G. This may help with providing some intuition for what potential shaping does, and how it behaves. Next, it is worth noting that PS_γ and $S'\mathrm{R}_\tau$ correspond to linear subspaces of $\mathcal{R}$. Specifically:

Proposition 33. *Let S be the set of all reward functions which can be expressed as*

$$R(s, a, s') = \gamma \cdot \Phi(s) - \Phi(s')$$

for some potential function Φ, and let Z be the set of all reward functions such that

$$\mathbb{E}_{S' \sim \tau(s,a)} \left[R(s, a, S') \right] = 0.$$

Then S and Z are linear subspaces of $\mathcal{R}$, where S is $|\mathcal{S}|$-dimensional and Z is $|\mathcal{S}||\mathcal{A}|(|\mathcal{S}| - 1)$-dimensional, and where $S \cap Z = R_0$.

Moreover, R_1 and R_2 differ by potential shaping if and only if $R_2 = R_1 + R'$ for some $R' \in S$, and R_1 and R_2 differ by S'-redistribution if and only if $R_2 = R_1 + R'$ for some $R' \in Z$.

Proof. To show that S and Z are linear subspaces of $\mathcal{R}$, we must show that $R_0 \in S$, that if $R_1, R_2 \in S$ then $R_1 + R_2 \in S$, and that if $R \in S$ then $c \cdot R \in S$ for all scalars c, and likewise for Z. Each of these properties are straightforward in both cases.

To see that S is $|\mathcal{S}|$-dimensional, for each state s, let R_s be the reward function in S which corresponds to the potential function Φ such that $\Phi(s) = 1$, and $\Phi(s') = 0$ for $s' \neq s$. Now the vectors $\{R_s : s \in \mathcal{S}\}$ form a basis for S. They are also linearly independent. To see this, recall Proposition 29. In particular, we have that $V^\pi(s) = -1$ for the reward function R_s, where π is any policy. However, for any reward function that can be expressed as a linear combination of reward functions in $\{R_s : s \in \mathcal{S}\} \setminus \{R_s\}$, we have that $V^\pi(s) = 0$. This means that $\{R_s : s \in \mathcal{S}\}$ is a minimal basis. Since $|\{R_s : s \in \mathcal{S}\}| = |\mathcal{S}|$, this means that S is $|\mathcal{S}|$-dimensional.

To see that Z is $|\mathcal{S}||\mathcal{A}|(|\mathcal{S}| - 1)$-dimensional, for each state-action pair s, a, pick a state s' such that $s' \in \mathrm{supp}(\tau(s, a))$. We can now set the reward of the transition

$\langle s, a, s'' \rangle$ freely for each s'' such that $s'' \neq s'$, given that $R(s, a, s')$ is selected to ensure that $\mathbb{E}_{S' \sim \tau(s,a)} [R(s, a, S')] = 0$. To see that $S \cap Z = R_0$, note that if $R \in Z$, then $V^\pi(s) = 0$ for every policy π and every state s. Then Proposition 29 implies that $\Phi(s) = 0$ for all s, and so $R = R_0$.

It can be shown from straightforward algebra that R_1 and R_2 differ by potential shaping if and only if $R_2 = R_1 + R'$ for some $R' \in S$, and likweise that R_1 and R_2 differ by S'-redistribution if and only if $R_2 = R_1 + R'$ for some $R' \in Z$. This completes the proof. □

Note that Proposition 33 implies that for any reward function R, the set of all reward functions that differ from R by potential shaping forms a $|\mathcal{S}|$-dimensional affine subspace of $\mathcal{R}$, and similarly for S'-redistribution. Moreover, since $S \cap Z = R_0$, we have that the set of all reward functions that differ from R by potential shaping and S'-redistribution forms a $(|\mathcal{S}||\mathcal{A}|(|\mathcal{S}| - 1) + |\mathcal{S}|)$-dimensional affine space.

4.2 Equivalent Reward Functions

In this section, we will introduce and study two important equivalence relations on $\mathcal{R}$. The first equivalence relation considers two reward functions to be equivalent if they have the same *ordering of policies*, and the second equivalence relation considers two reward functions to be equivalent if they have the same *optimal policies*. We will also characterise these equivalence relations in terms of reward transformations.

Given a discount γ and transition function τ, we say that $\text{ORD}_{\tau,\gamma}$ is the equivalence relation under which $R_1 \equiv_{\text{ORD}_{\tau,\gamma}} R_2$ if and only if R_1 and R_2 have the same *policy ordering* under τ and μ_0.[1] Moreover, we say that $\text{OPT}_{\tau,\gamma}$ is the equivalence relation under which $R_1 \equiv_{\text{OPT}_{\tau,\gamma}} R_2$ if and only if R_1 and R_2 have the same *optimal policies* under τ and μ_0. Note that if $R_1 \equiv_{\text{ORD}_{\tau,\gamma}} R_2$ then $R_1 \equiv_{\text{OPT}_{\tau,\gamma}} R_2$, but the converse does not hold — if two reward functions have the same policy ordering, then they have the same optimal policies, but they may

[1] Note that while the policy ordering of R may depend on the initial state distribution μ_0, we have that R_1 and R_2 have the same policy order for one μ_0 if and only if they have the same policy order for all μ_0, c.f. Theorem 40.

have the same optimal policies, without having the same policy ordering. This means that $\mathrm{ORD}_{\tau,\gamma}$ is stronger than $\mathrm{OPT}_{\tau,\gamma}$.

We will characterise these equivalence relations in terms of necessary and sufficient conditions on R_1 and R_2 (relative to a particular choice of transition function τ and discount factor γ). We first show that $\mathrm{OPT}_{\tau,\gamma}$ corresponds to optimality-preserving transformations:

Theorem 34. *$R_1 \equiv_{\mathrm{OPT}_{\tau,\gamma}} R_2$ if and only if $R_2 = t(R_1)$ for some $t \in \mathrm{OP}_{\tau,\gamma}$.*

Proof. Suppose R_1 and R_2 differ by an optimality-preserving transformation. Let Ψ be the corresponding value-bounding function, that is, a function $\Psi : \mathcal{S} \to \mathbb{R}$ satisfying, for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$,

$$\mathbb{E}_{S' \sim \tau(s,a)} \left[R_2(s, a, S') + \gamma \cdot \Psi(S') \right] \leq \Psi(s) \,,$$

with equality if and only if $a \in \mathrm{argmax} A_1^\star(s, \_)$. This gives us that

$$\Psi(s) = \max_{a \in \mathcal{A}} \left(\mathbb{E}_{S' \sim \tau(s,a)} \left[R_2(s, a, S') + \gamma \cdot \Psi(S') \right] \right) \,.$$

This recursive condition on Ψ is the Bellman optimality equation for the unique optimal value function $V_2^\star$ for R_2 (Equation 2.3). Therefore, $\Psi(s) = V_2^\star(s)$ for all $s \in \mathcal{S}$, and we can rewrite the above as

$$\mathbb{E}_{S' \sim \tau(s,a)} \left[R_2(s, a, S') + \gamma \cdot V_2^\star(S') \right] \leq V_2^\star(s) \,,$$

with equality if and only if $a \in \mathrm{argmax} A_1^\star(s, \_)$. This means that the actions which are optimal under R_1 are optimal under R_2, and vice versa, which in turn means that $R_1 \equiv_{\mathrm{OPT}_{\tau,\gamma}} R_2$.

Conversely, let R_1 and R_2 be any rewards such that $R_1 \equiv_{\mathrm{OPT}_{\tau,\gamma}} R_2$. This means that R_1 and R_2 share the same optimal actions. Let $V_2^\star$ and $A_2^\star$ denote the optimal value and advantage functions for R_2. The Bellman optimality equation for R_2 ensures that, for $s \in \mathcal{S}$,

$$V_2^\star(s) = \max_{a \in \mathcal{A}} \left(\mathbb{E}_{S' \sim \tau(s,a)} \left[R_2(s, a, S') + \gamma \cdot V_2^\star(S') \right] \right)$$

with the maximum attained precisely when $a \in \text{argmax}_{a \in \mathcal{A}}(A_2^\star(s,a))$. We thus have

$$\mathbb{E}_{S' \sim \tau(s,a)}\left[R_2(s,a,S') + \gamma \cdot V_2^\star(S')\right] \leq V_2^\star(s)$$

for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$, with equality if and only if $a \in \text{argmax}_{a \in \mathcal{A}}(A_2^\star(s,a))$. Note also that $\text{argmax}_{a \in \mathcal{A}}(A_2^\star(s,a)) = \text{argmax}_{a \in \mathcal{A}}(A_1^\star(s,a))$. This means that R_2 is produced from R_1 by an optimality-preserving transformation (with $\Psi(s) = V_2^\star(s)$), which completes the proof. □

Stated differently, Theorem 34 says that the MDPs $(\mathcal{S}, \mathcal{A}, \tau, \mu_0, R_1, \gamma)$ and $(\mathcal{S}, \mathcal{A}, \tau, \mu_0, R_2, \gamma)$ have the same optimal policies if and only if R_1 and R_2 differ by an optimality-preserving transformation.

Before we can characterise the equivalence relation $\text{ORD}_{\tau,\gamma}$, we must first derive a few lemmas about the topological structure of MDPs. Recall that the *occupancy measure* η^π of a policy π is the $(|\mathcal{S}||\mathcal{A}||\mathcal{S}|)$-dimensional vector in which the value of the (s,a,s')'th dimension is given by

$$\sum_{t=0}^{\infty} \gamma^t \mathbb{P}_{\xi \sim \pi}\left(S_t, A_t, S_{t+1} = s, a, s'\right).$$

Let $m_{\tau,\mu_0,\gamma} : \Pi \to \mathbb{R}^{|S||A||S|}$ be the map that sends each policy π to its occupancy measure, η^π. Recall also that $J(\pi) = \eta^\pi \cdot R$. This means that we can use $m_{\tau,\mu_0,\gamma}$ to decompose J into two separate steps, the first of which is independent of the reward function, and the second of which is linear. We will first show that $m_{\tau,\mu_0,\gamma}$ is a continuous function. Throughout this section, we will assume that Π is equipped with the topological structure that is induced by the L_2-norm, when each policy is represented as an $(|\mathcal{S}||\mathcal{A}|)$-dimensional vector (where the value of the (s,a)'th dimension is $\pi(a \mid s)$).

Lemma 35. *$m_{\tau,\mu_0,\gamma} : \Pi \to \mathbb{R}^{|S||A||S|}$ is a continuous function.*

Proof. Recall that a uniformly convergent series of continuous functions is continuous. Specifically, if X is a topological space and Y is a metric space, and $\{f_n : X \to Y\}_{n=1}^\infty$ is a sequence of functions that converge uniformly to a function $f : X \to Y$, and each function $f_i \in \{f_n\}_{n=1}^\infty$ is continuous, then f is continuous.

Moreover, $\{f_n\}_{n=1}^{\infty}$ converges uniformly to f if there for each ϵ exists an i such that for all $j \geq i$, $|f(x) - f_j(x)| < \epsilon$ for all $x \in X$. We will show that $m_{\tau,\mu_0,\gamma}$ can be expressed in this way.

Let $f_n : \Pi \to \mathbb{R}^{|S||A||S|}$ be the function that maps each policy π to its occupancy measure, when only the first n time steps are considered. That is, $f_n(\pi)$ is the vector in which the value of the (s, a, s')'th dimension is

$$\sum_{t=0}^{n} \gamma^t \mathbb{P}_{\xi\sim\pi}\left(S_t, A_t, S_{t+1} = s, a, s'\right).$$

Note that $\mathbb{P}_{\xi\sim\pi}\left(S_t, A_t, S_{t+1} = s, a, s'\right) \in [0, 1]$, and that $\gamma \in (0, 1)$. This means that $m_{\tau,\mu_0,\gamma}(\pi) \geq f_n(\pi)$, and that

$$m_{\tau,\mu_0,\gamma}(\pi) - f_n(\pi) = \sum_{t=n+1}^{\infty} \gamma^t \mathbb{P}_{\xi\sim\pi}\left(S_t, A_t, S_{t+1} = s, a, s'\right) \leq \left(\frac{\gamma^{n+1}}{1-\gamma}\right).$$

As n goes to ∞, we have that $\left(\frac{\gamma^{n+1}}{1-\gamma}\right)$ goes to 0. Thus, for any n that is sufficiently large, we have that $|m_{\tau,\mu_0,\gamma}(\pi) - f_n(\pi)| < \epsilon$. This means that $\{f_n\}_{n=1}^{\infty}$ converges uniformly to $m_{\tau,\mu_0,\gamma}$. Moreover, each function $f_i \in \{f_n\}_{n=1}^{\infty}$ is continuous, since it can be expressed as a finite sum of terms in which each term is given by a finite number of matrix multiplications. □

Next, let $\bar{\Pi} \subset \Pi$ be the set of all policies that visit each state with positive probability. We then have that:

Lemma 36. *$m_{\tau,\mu_0,\gamma}$ is injective on $\bar{\Pi}$.*

Proof. Suppose $m_{\tau,\mu_0,\gamma}(\pi) = m_{\tau,\mu_0,\gamma}(\pi')$ for some $\pi, \pi' \in \bar{\Pi}$. Next, given τ, μ_0, define w_π as

$$w_\pi(s) = \sum_{t=0}^{\infty} \gamma^t \mathbb{P}_{\xi\sim\pi}(S_t = s).$$

Note that if $m_{\tau,\mu_0,\gamma}(\pi) = m_{\tau,\mu_0,\gamma}(\pi')$ then $w_\pi = w_{\pi'}$, and moreover that

$$\sum_{s'\in\mathcal{S}} m_{\tau,\mu_0,\gamma}(\pi)[s, a, s'] = w_\pi(s)\pi(a \mid s).$$

This means that if $w_\pi(s) \neq 0$ for all s, which is the case for all $\pi \in \bar{\Pi}$, then we can express π as

$$\pi(a \mid s) = \frac{\sum_{s'\in\mathcal{S}} m_{\tau,\mu_0,\gamma}(\pi)[s, a, s']}{w_\pi(s)}.$$

This means that if $m_{\tau,\mu_0,\gamma}(\pi) = m_{\tau,\mu_0,\gamma}(\pi')$ for some $\pi, \pi' \in \bar{\Pi}$ then $\pi = \pi'$. □

Note that $m_{\tau,\mu_0,\gamma}$ is *not* injective on Π; if there is some state s that π reaches with probability 0, then we can alter the behaviour of π at s without changing $m_{\tau,\mu_0,\gamma}(\pi)$. Note also that Lemma 36 holds for all τ and μ_0, assuming that all states are reachable (which we assume throughout this work). We will also need the following lemma:

Lemma 37. $\mathrm{Im}(m_{\tau,\mu_0,\gamma})$ *is located in an affine space with no more than* $|\mathcal{S}|(|\mathcal{A}|-1)$ *dimensions.*

Proof. We wish to establish an *upper* bound on the number of linearly independent vectors in $\mathrm{Im}(m_{\tau,\mu_0,\gamma})$. We can do this by establishing a *lower* bound on the size of the space of all reward functions that share the same policy evaluation function, J. To see this, consider the fact that $J(\pi) = m_{\tau,\mu_0,\gamma}(\pi) \cdot R$, and note that R is an $|\mathcal{S}||\mathcal{A}||\mathcal{S}|$-dimensional vector. Let R_1 be a reward function, and let X be the space of all reward functions R_2 such that $R_1 \cdot \eta = R_2 \cdot \eta$ for all $\eta \in \mathrm{Im}(m_{\tau,\mu_0,\gamma})$. It is then a straightforward consequence of linear algebra that if $\mathrm{Im}(m_{\tau,\mu_0,\gamma})$ contains n linearly independent vectors, then X forms an affine space with $|\mathcal{S}||\mathcal{A}||\mathcal{S}| - n$ dimensions. We can thus obtain an upper bound on the number of linearly independent vectors in $\mathrm{Im}(m_{\tau,\mu_0,\gamma})$ from a lower bound on the dimensionality of X.

Next, recall that if R_1 and R_2 differ by potential shaping with Φ, and $\mathbb{E}_{S_0\sim\mu_0}[\Phi(S_0)] = 0$, then $J_1(\pi) = J_2(\pi)$ for all π (Proposition 29). Also recall that if R_1 and R_2 differ by S'-redistribution, then $J_1(\pi) = J_2(\pi)$ for all π. This means that for any R_1, we have that X contains all reward functions R_2 that differ from R_1 by S'-redistribution and potential shaping with a potential function Φ such that $\mathbb{E}_{S_0\sim\mu_0}[\Phi(S_0)] = 0$. The space of all such reward vectors is an affine space with $|\mathcal{S}||\mathcal{A}|(|\mathcal{S}|-1) + |\mathcal{S}| - 1$ dimensions (Proposition 33). This means that $\mathrm{Im}(m_{\tau,\mu_0,\gamma})$ contains at most $|\mathcal{S}|(|\mathcal{A}|-1)+1$ linearly independent vectors.

Next, note that there is no π such that $m_{\tau,\mu_0,\gamma}(\pi)$ is the zero vector. In fact, $\sum m_{\tau,\mu_0,\gamma}(\pi) = 1/(1-\gamma)$ for all π. This means that the smallest affine space which contains $\mathrm{Im}(m_{\tau,\mu_0,\gamma})$ does not contain the origin. Therefore, $\mathrm{Im}(m_{\tau,\mu_0,\gamma})$ is located in an affine space with no more than $|\mathcal{S}|(|\mathcal{A}|-1)$ dimensions. □

For the next lemma, let $\Pi^+ \subset \Pi$ be the set of all policies that take all actions with positive probability in each state. Note that $\Pi^+ \subset \bar{\Pi}$ (i.e., a policy that takes every action with positive probability in each state visits every state with positive probability), since we assume that all states are reachable under μ_0 and τ. We then have that:

Lemma 38. $\mathrm{Im}(m_{\tau,\mu_0,\gamma})$ *is located in an affine space with* $|\mathcal{S}|(|\mathcal{A}|-1)$ *dimensions, in which* $m_{\tau,\mu_0,\gamma}(\Pi^+)$ *is an open set, and* $m_{\tau,\mu_0,\gamma}$ *is a homeomorphism between* Π^+ *and* $m_{\tau,\mu_0,\gamma}(\Pi^+)$.

Proof. By the Invariance of Domain theorem, if

1. U is an open subset of $\mathbb{R}^n$, and
2. $f : U \to \mathbb{R}^n$ is an injective continuous map,

then $f(U)$ is open in $\mathbb{R}^n$, and f is a homeomorphism between U and $f(U)$. We will show that m and Π^+ satisfy the requirements of this theorem.

We begin by noting that Π can be represented as a set of points in $\mathbb{R}^{|\mathcal{S}|(|\mathcal{A}|-1)}$. We do this by considering each policy π as a vector $\vec{\pi}$ of length $|\mathcal{S}||\mathcal{A}|$, where $\vec{\pi}[s,a] = \pi(a \mid s)$. Moreover, since $\sum_{a \in A} \pi(a \mid s) = 1$ for all s, we can remove $|\mathcal{A}|$ dimensions, and embed Π in $\mathbb{R}^{|\mathcal{S}|(|\mathcal{A}|-1)}$.

Π^+ is an open set in $\mathbb{R}^{|\mathcal{S}|(|\mathcal{A}|-1)}$. By Lemma 37, we have that $m_{\tau,\mu_0,\gamma}$ is a mapping from Π^+ to an affine space with no more than $|\mathcal{S}|(|\mathcal{A}|-1)$ dimensions. By Lemma 36, we have that $m_{\tau,\mu_0,\gamma}$ is injective on Π^+. Finally, by Lemma 35, we have that $m_{\tau,\mu_0,\gamma}$ is continuous. We can therefore apply the Invariance of Domain theorem, and conclude that $m_{\tau,\mu_0,\gamma}(\Pi^+)$ is open in this $|\mathcal{S}|(|\mathcal{A}|-1)$-dimensional affine space, and that $m_{\tau,\mu_0,\gamma}$ is a homeomorphism between Π^+ and $m_{\tau,\mu_0,\gamma}(\Pi^+)$. □

Note that lemma 38 holds for all τ and μ_0 (for which all states are reachable). Using these results, we can now state necessary and sufficient conditions that describe when $J_1(\pi) = J_2(\pi)$ for all policies π:

Lemma 39. $J_1 = J_2$ *if and only if* R_1 *and* R_2 *differ by* S'*-redistribution and potential shaping with a potential* Φ *such that* $\mathbb{E}_{S_0 \sim \mu_0}[\Phi(S_0)] = 0$.

Proof. For the first direction, suppose R_1 and R_2 differ by S'-redistribution and potential shaping with a potential Φ such that $\mathbb{E}_{S_0 \sim \mu_0}[\Phi(S_0)] = 0$. Then $V_2^\pi(s) = V_1^\pi(s) - \Phi(s)$, as per Proposition 29. Hence $J_1(\pi) = J_2(\pi) - \mathbb{E}_{s_0 \sim \mu_0}[\Phi(s_0)] = J_2(\pi)$, and so we have proven the first direction.

For the other direction, first recall that $J(\pi) = m_{\tau,\mu_0,\gamma}(\pi) \cdot R$. Next, Lemma 38 implies that $\mathrm{Im}(m_{\tau,\mu_0,\gamma})$ contains $|S|(|A|-1)+1$ linearly independent vectors. It is then a straightforward fact of linear algebra that, for any reward function R_1, the space X of all reward functions R_2 such that $R_1 \cdot \eta = R_2 \cdot \eta$ for all $\eta \in \mathrm{Im}(m_{\tau,\mu_0,\gamma})$, forms an affine space with $|\mathcal{S}||\mathcal{A}||\mathcal{S}| - (|S|(|A|-1)+1) = |\mathcal{S}||\mathcal{A}|(|\mathcal{S}|-1) + |\mathcal{S}| - 1$ dimensions.

We have already shown that $J_1(\pi) = J_2(\pi)$ for all policies π if R_1 and R_2 differ by S'-redistribution and potential shaping with a potential Φ such that $\mathbb{E}_{S_0 \sim \mu_0}[\Phi(S_0)] = 0$. Next, given R_1, the space of all reward functions R_2 such that R_1 and R_2 differ by S'-redistribution and potential shaping with a potential Φ such that $\mathbb{E}_{S_0 \sim \mu_0}[\Phi(S_0)] = 0$, forms an affine space with $|\mathcal{S}||\mathcal{A}|(|\mathcal{S}|-1) + |\mathcal{S}| - 1$ dimensions (Proposition 33). Since this space is contained in X, and since they have the same number of dimensions, they must be one and the same. Therefore, if $J_1 = J_2$, then it must be the case that R_1 and R_2 differ by S'-redistribution and potential shaping with a potential Φ such that $\mathbb{E}_{S_0 \sim \mu_0}[\Phi(S_0)] = 0$. We have thus proven the other direction, which completes the proof. □

Using these results, we can now derive necessary and sufficient conditions for two reward functions R_1, R_2 to have the same ordering of policies (relative to a particular choice of τ and μ_0):

Theorem 40. *For all $R_1, R_2 \in \mathcal{R}$, all τ and all γ, we have that $R_1 \equiv_{\mathrm{ORD}_{\tau,\gamma}} R_2$ if and only if $R_2 = t(R_1)$ for some $t \in S'\mathrm{R}_\tau \bigodot \mathrm{PS}_\gamma \bigodot \mathrm{LS}$.*

Proof. The first direction is straightforward. First, if $R_1 = t(R_2)$ for some $t \in S'\mathrm{R}_\tau$ then $J_1 = J_2$. Next, if $R_1 = t(R_2)$ for some $t \in \mathrm{PS}_\gamma$ then $J_1 = J_2 - \mathbb{E}_{S_0 \sim \mu_0}[\Phi_t(S_0)]$ (Proposition 29). Finally, if $R_1 = t(R_2)$ for some $t \in \mathrm{LS}$ then $J_1 = c \cdot J_2$ for some $c \in \mathbb{R}^+$. Hence if $R_1 = t(R_2)$ for some $t \in S'\mathrm{R}_\tau \bigodot \mathrm{PS}_\gamma \bigodot \mathrm{LS}$, then $J_1 = a \cdot J_2 + b$

for some $a \in \mathbb{R}^+, b \in \mathbb{R}$. This means that J_1 and J_2 differ by a strictly monotonic transformation, and so R_1 and R_2 have the same ordering of policies.

For the other direction, first note that R_1 and R_2 have the same ordering of policies only if J_1 is a monotonic transformation of J_2. Moreover, since $J(\pi) = m_{\tau,\mu_0,\gamma}(\pi) \cdot R$, we have that all possible monotonic transformations of J are affine. Hence R_1 and R_2 have the same ordering of policies only if $J_1 = a \cdot J_2 + b$ for some $a \in \mathbb{R}^+, b \in \mathbb{R}$.

Now suppose $J_1 = a \cdot J_2 + b$ for some $a \in \mathbb{R}^+, b \in \mathbb{R}$. Consider the reward function R_3 given by first scaling R_2 by a, and then shaping the resulting reward with the potential function Φ that is equal to $-b$ for all initial states, and equal to 0 elsewhere. Now $J_3 = J_1$, so (by Lemma 39) we can produce R_1 from R_3 by S'-redistribution and potential shaping with some potential function Φ such that $\mathbb{E}_{S_0 \sim \mu_0}[\Phi(S_0)] = 0$. By composing these transformations with the transformation that produced R_3 from R_2, we obtain a $t \in S'\mathrm{R}_\tau \bigodot \mathrm{PS}_\gamma \bigodot \mathrm{LS}$ such that $R_1 = t(R_2)$. Hence if R_1 and R_2 have the same ordering of policies then $R_1 = t(R_2)$ for some $t \in S'\mathrm{R}_\tau \bigodot \mathrm{PS}_\gamma \bigodot \mathrm{LS}$. We have thus proven both directions. □

Stated differently, Theorem 40 says that the MDPs $(\mathcal{S}, \mathcal{A}, \tau, \mu_0, R_1, \gamma)$ and $(\mathcal{S}, \mathcal{A}, \tau, \mu_0, R_2, \gamma)$ have the same ordering of policies if and only if R_1 and R_2 differ by potential shaping, positive linear scaling, and S'-redistribution. This result will be very important for our analysis.

In Section 8.5, we discuss the question of how to define and characterise even stronger equivalence relations on $\mathcal{R}$.

4.3 Unhackable Reward Functions

Recall that we require a method for comparing reward functions in order to formalise what it means for the learnt reward function R_H to be "close" to the true reward function, $R^\star$. So far, we have formalised this in terms of *equivalence relations* and *pseudometrics*. In this section, we briefly explore another option. Specifically, we will define the notion of *unhackability*, which is a non-transitive relationship

between reward functions, and argue that it would be sufficient for R_H and $R^\star$ to be unhackable to ensure that R_H is "close" to $R^\star$. However, we will then show that no reward functions are unhackable relative to each other, except in trivial cases. We thus conclude that the notion of unhackability, though initially promising, cannot be used in our later analysis.

Our motivation for considering the notion of "unhackability" is the observation that a misspecified proxy reward in practice is problematic primarily when it is possible to increase the proxy reward while decreasing the "intended" reward. For example, a chess-playing agent trained to maximise material gain may be incentivised to pursue a strategy that has a high probability of losing the game, provided that the strategy leads to a sufficiently large material gain during the match. Note that what is problematic about this reward is not only that it may fail to induce good chess-playing behaviour, but also that a good chess-playing policy may get *worse* as a result of being optimised for this reward. This leads us to introduce the following definition:

Definition 41. Given a transition function τ and a discount factor γ, we say that two rewards R_1 and R_2 are *unhackable* if there are no policies π_1, π_2 such that

$$J_1(\pi_1) > J_1(\pi_2) \text{ but } J_2(\pi_1) < J_2(\pi_2).$$

Intuitively, R_1 and R_2 are unhackable if there is no way to *increase* R_1 while *decreasing* R_2. Note that this is not equivalent to saying that R_1 and R_2 have the same policy ordering, because Definition 41 permits there to be policies π_1, π_2 such that $J_1(\pi_1) > J_1(\pi_2)$ but $J_2(\pi_1) = J_2(\pi_2)$, etc. For example, we could define an analogous notion for real-valued functions, and say that $f, g : \mathbb{R} \to \mathbb{R}$ are unhackable if there are no $x, y \in \mathbb{R}$ such that $f(x) > f(y)$ but $g(x) < g(y)$. In that case, we would find that the Heaviside step function and the tanh function are unhackable, even though they induce different orderings over $\mathbb{R}$. If R_1 and R_2 are unhackable in this sense, then it seems reasonable to say that R_1 is similar to

(or, at least, safe from the point of view of) R_2, because an agent that maximises R_1 is never *actively incentivised* to decrease R_2.[2]

Now that we have this definition, we can immediately make a few preliminary observations:

1. The trivial reward function R_0 is unhackable with respect to any other reward function. This tells us that Definition 41 can be satisfied, and that it does not imply that R_1 and R_2 must have the same policy ordering.

2. Definition 41 is not transitive. In particular, for any two reward functions R_1 and R_2, we have that R_1 and R_0 are unhackable, and that R_0 and R_2 are unhackable. However, it is not the case that any two reward functions are unhackable.[3] This means that Definition 41 is not an equivalence relation.

In light of this, it seems like Definition 41 may be a promising candidate for an alternative way to formalise what it should mean for the learnt reward R_H to be "similar" to the true reward $R^\star$. Unfortunately, our next result shows that there are no interesting ways to satisfy Definition 41:

Theorem 42. *For any τ and any γ, if R_1 and R_2 are unhackable, then either $R_1 \equiv_{\mathrm{ORD}_{\tau,\gamma}} R_2$, or at least one of R_1 and R_2 is trivial.*

Proof. Let R_1 and R_2 be unhackable and non-trivial. We will begin by showing that if $J_1(\pi_1) = J_1(\pi_2)$ then $J_2(\pi_1) = J_2(\pi_2)$. First recall that $J_1(\pi)$ and $J_2(\pi)$ can be expressed as $\eta^\pi \cdot R_1$ and $\eta^\pi \cdot R_2$, where $\eta^\pi = m_{\tau,\mu_0,\gamma}(\pi)$ is the occupancy measure of π. Also recall that the set of all occupancy measures, Ω, lies in an affine space Z of $|S|(|A|-1)$ dimensions (Lemma 37), and that multiplication by R_1 or R_2 induces two linear functions over Z. Let π_1, π_2 be any two policies such that

[2]Note, however, that it is still possible for an agent to decrease R_2 while optimising R_1, because there could be policies π_1, π_2 such that $J_1(\pi_1) = J_1(\pi_2)$ but $J_2(\pi_1) < J_2(\pi_2)$. Thus, an agent which optimises R_1 could start with a policy π_2 and replace it with π_1 — this is equivalent according to R_1, but worse according to R_2. However, such reductions in performance (according to R_2) could only happen as a result of random drift (according to R_1), which seems much less concerning than the case where R_1 can actively incentivise the agent to reduce R_2.

[3]For example, if R is non-trivial, then R and $-R$ are not unhackable.

$J_1(\pi_1) = J_1(\pi_2)$, and let $\bar{\pi}$ be an arbitrary policy in $\bar{\Pi}$ (where $\bar{\Pi}$ is the set of all policies that visit all states with positive probability).

Let y_1 and y_2 be the vectors such that $\eta^{\pi_1} = \eta^{\bar{\pi}} + y_1$ and $\eta^{\pi_2} = \eta^{\bar{\pi}} + y_2$. Since $m_{\tau,\mu_0,\gamma}(\bar{\Pi})$ is open in Z (as per Lemma 38), since $\eta^{\pi_1}, \eta^{\pi_2} \in Z$, and since Z is affine, we have that there is a $\delta > 0$ such that $\eta^{\bar{\pi}} + \delta y_1$ and $\eta^{\bar{\pi}} + \delta y_2$ both are contained in $m_{\tau,\mu_0,\gamma}(\bar{\Pi})$. Moreover, recall that $J_1(\pi_1) = R_1 \cdot \eta^{\pi_1} = R_1 \cdot (\eta^{\bar{\pi}} + y_1)$ and $J_1(\pi_2) = R_1 \cdot \eta^{\pi_2} = R_1 \cdot (\eta^{\bar{\pi}} + y_2)$. This gives us that

$$
\begin{aligned}
J_1(\pi_1) = J_1(\pi_2) &\implies \\
R_1 \cdot (\eta^{\bar{\pi}} + y_1) = R_1 \cdot (\eta^{\bar{\pi}} + y_2) &\implies \\
R_1 \cdot \eta^{\bar{\pi}} + R_1 \cdot y_1 = R_1 \cdot \eta^{\bar{\pi}} + R_1 \cdot y_2 &\implies \\
R_1 \cdot y_1 = R_1 \cdot y_2 &\implies \\
R_1 \cdot \delta y_1 = R_1 \cdot \delta y_2 &\implies \\
R_1 \cdot (\eta^{\bar{\pi}} + \delta y_1) = R_1 \cdot (\eta^{\bar{\pi}} + \delta y_2)&
\end{aligned}
$$

Thus $J_1(\pi_1') = J_1(\pi_2')$, where π_1' and π_2' are the policies whose occupancy measures are $(\eta^{\bar{\pi}} + \delta y_1)$ and $(\eta^{\bar{\pi}} + \delta y_2)$ respectively.

Since neither R_1 or R_2 is trivial, there exists two hyperplanes H_1 and H_2 such that $(\eta^{\bar{\pi}} + \delta y_1)$ is located on both H_1 and H_2, and such that $R_1 \cdot x$ is equal to $R_1 \cdot (\eta^{\bar{\pi}} + \delta y_1)$ for all points $x \in H_1$, and $R_2 \cdot x$ is equal to $R_2 \cdot (\eta^{\bar{\pi}} + \delta y_1)$ for all points $x \in H_2$. Now, if it were the case that $H_1 \neq H_2$, then it would be the case that there are points in $\text{Im}(\tilde{\Pi})$ that are located above H_1 but below H_2, or vice versa. This would imply that there are policies $\tilde{\pi_1}, \tilde{\pi_2} \in \tilde{\Pi}$ such that $J_1(\tilde{\pi_1}) > J_1(\tilde{\pi_2})$ but $J_2(\tilde{\pi_1}) < J_2(\tilde{\pi_2})$, or vice versa. Since R_1 and R_2 are unhackable, this cannot happen. That means that $H_1 = H_2$, which thus implies that $(\eta^{\bar{\pi}} + \delta y_2)$ is located on H_2, and hence that $R_2 \cdot (\eta^{\bar{\pi}} + \delta y_1) = R_2 \cdot (\eta^{\bar{\pi}} + \delta y_2)$. By the properties of linear functions, this implies that $R_2 \cdot (\eta^{\bar{\pi}} + y_1) = R_2 \cdot (\eta^{\bar{\pi}} + y_2)$, which in turn implies that $J_2(\pi_1) = J_2(\pi_2)$.

Since π_1 and π_2 were picked arbitrarily, this implies that if $J_1(\pi_1) = J_1(\pi_2)$ then $J_2(\pi_1) = J_2(\pi_2)$. By symmetry, we thus have that for all π_1, π_2,

$$J_1(\pi_1) = J_1(\pi_2) \iff J_2(\pi_1) = J_2(\pi_2).$$

Since R_1 and R_2 are unhackable, this then implies that $R_1 \equiv_{\text{ORD}_{\tau,\gamma}} R_2$. □

Of course, the case where R_1 or R_2 is trivial is uninteresting. Thus, Definition 41 effectively collapses to saying that R_1 and R_2 have the same policy order. This, in turn, means that it cannot be used to reveal any interesting structure not already revealed by our analysis using the equivalence relation $\equiv_{\text{ORD}_{\tau,\gamma}}$.[4]

4.4 STARC Metrics

In this section, we introduce a family of *pseudometrics* on $\mathcal{R}$, which we can use to get a fine-grained quantification of the difference between any two reward functions. First, we note that it is not straightforward to quantify the difference between reward functions in an informative way. A simple method might be to measure their L_2-distance. However, this is unsatisfactory, because two reward functions can have a large L_2-distance, even if they induce the *same* ordering of policies, or a small L_2-distance, even if they induce the *opposite* ordering of policies. For example, given an arbitrary reward function R and an arbitrary constant c, we have that R and $c \cdot R$ have the same ordering of policies, even though their L_2-distance may be arbitrarily large. Similarly, for any ϵ, we have that $\epsilon \cdot R$ and $-\epsilon \cdot R$ have the opposite ordering of policies, unless R is trivial, even though their L_2-distance may be arbitrarily small. Constructing a pseudometric on $\mathcal{R}$ which provides an informative quantification of the difference between two reward functions will therefore require more care.

Before proceeding, we should consider what properties a function $d : \mathcal{R} \times \mathcal{R} \to \mathbb{R}$ needs to have, in order to provide a useful way of quantifying the difference between reward functions. First of all, it would certainly be desirable for d to be a *pseudometric*, since pseudometrics provide a well-defined notion of "distance" that can be used in mathematical analysis. Moreover, it seems reasonable to permit d to be a pseudometric, rather than to require d to be a (proper) metric,

[4]It is worth noting that Theorem 42 relies on specific properties of MDPs, policies, and reward functions. For example, in the case of real-valued functions, the analogous statement does not hold — as already noted, the Heaviside step function and the tanh function are unhackable and non-trivial (if these definitions are modified to fit real-valued functions), but they do induce different orderings over $\mathbb{R}$.

because we may want to consider distinct reward functions to be equivalent. For example, if R_1 and R_2 have the same ordering of policies, then it would be natural to consider their distance to be 0.

Moreover, it would be highly desirable for d to induce an upper bound on worst-case regret. Specifically, we want it to be the case that if $d(R_1, R_2)$ is small, then the impact of optimising R_2 instead of R_1 should also be small. When a pseudometric has this property, we say that it is *sound*:

Definition 43. A pseudometric d on $\mathcal{R}$ is *sound* if there exists a positive constant U, such that for any reward functions R_1 and R_2, if two policies π_1 and π_2 satisfy that $J_2(\pi_2) \geq J_2(\pi_1)$, then

$$J_1(\pi_1) - J_1(\pi_2) \leq U \cdot (\max_{\pi} J_1(\pi) - \min_{\pi} J_1(\pi)) \cdot d(R_1, R_2).$$

Let us unpack this definition. $J_1(\pi_1) - J_1(\pi_2)$ is the regret, as measured by R_1, of using policy π_2 instead of π_1. Division by $\max_\pi J_1(\pi) - \min_\pi J_1(\pi)$ normalises this quantity based on the total range of R_1 (though the term is put on the right-hand side of the inequality, instead of being used as a denominator, in order to avoid division by zero when R_1 is trivial). The condition that $J_2(\pi_2) \geq J_2(\pi_1)$ says that R_2 prefers π_2 over π_1. Taken together, this means that a pseudometric d on $\mathcal{R}$ is sound if $d(R_1, R_2)$ gives an upper bound on the maximal regret that could be incurred under R_1 if an arbitrary policy π_1 is optimised to another policy π_2 according to R_2. It is worth noting that this includes the special case when π_1 is optimal under R_1 and π_2 is optimal under R_2. It is also worth noting that Definition 43 implicitly is given relative to a particular choice of τ and γ (via J_1 and J_2).

Moreover, it would also be preferable for d to induce a *lower* bound on worst-case regret. It may not be immediately obvious why this property is desirable. To see why this is the case, note that if a pseudometric d on $\mathcal{R}$ does not induce a lower bound on worst-case regret, then there are reward functions that have a *low* worst-case regret, but a *large* distance under d. This would in turn mean that d is not tight, and that it should be possible to improve upon it. In other words, if we want a small distance under d to be both sufficient *and necessary* for low worst-case

regret, then d must induce both an upper *and a lower* bound on worst-case regret. As such, we also introduce the following definition:

Definition 44. A pseudometric d on $\mathcal{R}$ is *complete* if there exists a positive constant L, such that for any reward functions R_1 and R_2, there exists two policies π_1 and π_2 such that $J_2(\pi_2) \geq J_2(\pi_1)$ and

$$J_1(\pi_1) - J_1(\pi_2) \geq L \cdot (\max_{\pi} J_1(\pi) - \min_{\pi} J_1(\pi)) \cdot d(R_1, R_2),$$

and moreover, if both R_1 and R_2 are trivial, then $d(R_1, R_2) = 0$.

The last condition is included to rule out certain pathological edge-cases. Intuitively, if d is sound, then a small d is *sufficient* for low regret, and if d is complete, then a small d is *necessary* for low regret. Soundness implies the absence of false positives, and completeness the absence of false negatives. Soundness and completeness also implies the following property:

Proposition 45. *If a pseudometric d on $\mathcal{R}$ is both sound and complete, then $d(R_1, R_2) = 0$ if and only if $R_1 \equiv_{\mathrm{ORD}_{\tau,\gamma}} R_2$.*

Proof. For the first direction, assume that R_1 and R_2 have the same ordering of policies. If both R_1 and R_2 are trivial, then the definition of completeness directly implies that $d(R_1, R_2) = 0$. Next, assume that R_1 and R_2 are not trivial, and assume for contradiction that $d(R_1, R_2) > 0$. Since d is complete, there exists two policies π_1 and π_2 such that $J_2(\pi_2) \geq J_2(\pi_1)$ and

$$J_1(\pi_1) - J_1(\pi_2) \geq L \cdot (\max_{\pi} J_1(\pi) - \min_{\pi} J_1(\pi)) \cdot d(R_1, R_2)$$

for some $L > 0$. Since R_1 is non-trivial, we have that $(\max_\pi J_1(\pi) - \min_\pi J_1(\pi) > 0$, which means that the right-hand side of this expression is positive. This implies that there are policies π_1 and π_2 such that $J_2(\pi_2) \geq J_2(\pi_1)$ but $J_1(\pi_2) < J_1(\pi_1)$, which is a contradiction, since R_1 and R_2 have the same policy order. Thus, if R_1 and R_2 have the same policy order, then $d(R_1, R_2) = 0$.

For the other direction, assume that $d(R_1, R_2) = 0$. Since d is sound, we have that there exists a positive constant U such that if two policies π_1 and π_2 satisfy that $J_2(\pi_2) \geq J_2(\pi_1)$, then

$$J_1(\pi_1) - J_1(\pi_2) \leq U \cdot (\max_{\pi} J_1(\pi) - \min_{\pi} J_1(\pi)) \cdot d(R_1, R_2).$$

Since $d(R_1, R_2) = 0$, this means that $J_1(\pi_1) - J_1(\pi_2) \leq 0$, which means that $J_1(\pi_2) \geq J_1(\pi_1)$. Since π_1 and π_2 were chosen arbitrarily, this means that if $J_2(\pi_2) \geq J_2(\pi_1)$, then $J_1(\pi_2) \geq J_1(\pi_1)$. As such, R_1 and R_2 have the same policy order. □

In other words, a pseudometric that is sound and complete must consider two reward functions to be equivalent exactly when they induce the same ordering of policies. Next, it is worth noting that if two pseudometrics d_1, d_2 on $\mathcal{R}$ are both sound and complete, then d_1 and d_2 are bilipschitz equivalent. This means that if there is a pseudometric on $\mathcal{R}$ that is both sound and complete, then this pseudometric is unique up to bilipschitz equivalence:

Proposition 46. *Any pseudometrics on $\mathcal{R}$ that are both sound and complete are bilipschitz equivalent.*

Proof. Assume that d_1 and d_2 are pseudometrics on $\mathcal{R}$ that are both sound and complete. Since d_1 is complete, we have that

$$L_1 \cdot d_1(R_1, R_2) \cdot (\max_{\pi} J_1(\pi) - \min_{\pi} J_1(\pi)) \leq \max_{\pi_1, \pi_2 : J_2(\pi_2) \geq J_2(\pi_1)} J_1(\pi_1) - J_1(\pi_2).$$

Similarly, since d_2 is sound, we also have that

$$\max_{\pi_1, \pi_2 : J_2(\pi_2) \geq J_2(\pi_1)} J_1(\pi_1) - J_1(\pi_2) \leq U_2 \cdot d_2(R_1, R_2) \cdot (\max_{\pi} J_1(\pi) - \min_{\pi} J_1(\pi)).$$

This implies that

$$\begin{aligned} &L_1 \cdot d_1(R_1, R_2) \cdot (\max_{\pi} J_1(\pi) - \min_{\pi} J_1(\pi)) \\ \leq &U_2 \cdot d_2(R_1, R_2) \cdot (\max_{\pi} J_1(\pi) - \min_{\pi} J_1(\pi)). \end{aligned}$$

First suppose that $(\max_\pi J_1(\pi) - \min_\pi J_1(\pi)) > 0$. We can then divide both sides, and obtain that

$$d_1(R_1, R_2) \leq \left(\frac{U_2}{L_1}\right) d_2(R_1, R_2).$$

Similarly, we also have that

$$\left(\frac{L_2}{U_1}\right) d_2(R_1, R_2) \leq d_1(R_1, R_2).$$

This means that we have constants $\left(\frac{U_2}{L_1}\right)$ and $\left(\frac{L_2}{U_1}\right)$ not depending on R_1 or R_2, such that

$$\left(\frac{L_2}{U_1}\right) d_2(R_1, R_2) \leq d_1(R_1, R_2) \leq \left(\frac{U_2}{L_1}\right) d_2(R_1, R_2)$$

for all R_1 and R_2 such that $(\max_\pi J_1(\pi) - \min_\pi J_1(\pi)) > 0$.

Next let $(\max_\pi J_1(\pi) - \min_\pi J_1(\pi)) = 0$ but $(\max_\pi J_2(\pi) - \min_\pi J_2(\pi)) > 0$. Since d_1 and d_2 are pseudometrics, we have that $d_1(R_1, R_2) = d_1(R_2, R_1)$ and $d_2(R_1, R_2) = d_2(R_2, R_1)$. Therefore, $\left(\frac{L_2}{U_1}\right) d_2(R_1, R_2) \leq d_1(R_1, R_2) \leq \left(\frac{U_2}{L_1}\right) d_2(R_1, R_2)$ in this case as well, as already shown above.

Finally, let $(\max_\pi J_1(\pi) - \min_\pi J_1(\pi)) = 0$ and $(\max_\pi J_2(\pi) - \min_\pi J_2(\pi)) = 0$. In this case, R_1 and R_2 induce the same policy order. This in turn means that $d_1(R_1, R_2) = d_2(R_1, R_2) = 0$, and so

$$\left(\frac{L_2}{U_1}\right) d_2(R_1, R_2) \leq d_1(R_1, R_2) \leq \left(\frac{U_2}{L_1}\right) d_2(R_1, R_2)$$

in this case as well. This completes the proof. □

We will next derive a family of pseudometrics on $\mathcal{R}$, which we refer to as *STAndardised Reward Comparison (STARC) metrics*, and show that these pseudometrics are both sound and complete. This means that all pseudometrics in this family induce both an upper and a lower bound on worst-case regret, and that any other pseudometric with this property must be bilipschitz equivalent to our metrics. As such, STARC metrics can be considered to be canonical, in a certain sense.

STARC metrics are computed in several steps, where the first steps collapse certain equivalence classes in $\mathcal{R}$ to a single representative, and the last step measures a distance. Recall that Proposition 45 says that if d is both sound and complete,

then $d(R_1, R_2) = 0$ if and only if R_1 and R_2 have the same policy order. Moreover, also recall that Theorem 40 says that R_1 and R_2 have the same policy order if and only if R_1 and R_2 differ by potential shaping, S'-redistribution, and positive linear scaling. This implies that if d is sound and complete, then $d(R_1, R_2) = 0$ if and only if R_1 and R_2 differ by potential shaping, S'-redistribution, and positive linear scaling. Our metrics are therefore computed by first standardising the reward functions to ensure that rewards are considered to be equivalent when they differ by these transformations. After this, the distance can be measured.

The first step standardises potential shaping and S'-redistribution. These transformations can be characterised in terms of linear subspaces of $\mathcal{R}$ (Proposition 33), which means that this standardisation can be achieved by a linear transformation:

Definition 47. A function $c : \mathcal{R} \to \mathcal{R}$ is a *canonicalisation function* if c is linear, $c(R)$ and R differ by potential shaping and S'-redistribution, and $c(R_1) = c(R_2)$ if and only if R_1 and R_2 only differ by potential shaping and S'-redistribution.

A canonicalisation function is a quotient map for the subspace of $\mathcal{R}$ that is given by potential shaping and S'-redistribution. Note that we require c to be linear. We will later provide examples of canonicalisation functions. Let us next introduce the functions that we use to compute a distance:

Definition 48. A metric $m : \mathcal{R} \times \mathcal{R} \to \mathbb{R}$ is *admissible* if there exists a norm p and two (positive) constants u, ℓ such that $\ell \cdot p(x, y) \leq m(x, y) \leq u \cdot p(x, y)$ for all $x, y \in \mathcal{R}$.

A metric is admissible if it is bilipschitz equivalent to a norm. Any norm is an admissible metric, though there are admissible metrics which are not norms.[5] Recall also that all norms are bilipschitz equivalent on any finite-dimensional vector space. This means that if m satisfies Definition 48 for one norm, then it satisfies it for all norms. Given these two components, we can now define our class of reward metrics:

[5]For example, the unit ball of m does not have to be convex, or symmetric around the origin, etc.

Definition 49. A function $d : \mathcal{R} \times \mathcal{R} \to \mathbb{R}$ is a *STARC metric* (STAndardised Reward Comparison) if there is a canonicalisation function c, a function n that is a norm on $\text{Im}(c)$, and a metric m that is admissible on $\text{Im}(s)$, such that $d(R_1, R_2) = m(s(R_1), s(R_2))$, where $s(R) = c(R)/n(c(R))$ when $n(c(R)) \neq 0$, and $c(R)$ otherwise.

Intuitively speaking, c ensures that all reward functions which differ by potential shaping and S'-redistribution are considered to be equivalent, and division by n ensures that positive linear scaling is ignored as well. Note that if $n(c(R)) = 0$, then $c(R) = R_0$. Note also that $\text{Im}(c)$ is the image of c, if c is applied to the entirety of $\mathcal{R}$, and similarly for $\text{Im}(s)$. If n is a norm on $\mathcal{R}$, then n is also a norm on $\text{Im}(c)$, but there are functions which are norms on $\text{Im}(c)$ but not on $\mathcal{R}$ (c.f. Proposition 55), and similarly for $\text{Im}(s)$.

STARC metrics have a number of important properties. We first note that STARC metrics indeed are pseudometrics on $\mathcal{R}$, which means that they give us a well-defined notion of a "distance" between reward functions:

Proposition 50. *All STARC metrics are pseudometrics on $\mathcal{R}$.*

Proof. To show that d is a pseudometric, we must show that

1. $d(R, R) = 0$
2. $d(R_1, R_2) = d(R_2, R_1)$
3. $d(R_1, R_3) \leq d(R_1, R_2) + d(R_2, R_3)$

1 follows from the fact that m is a metric, and 2 follows directly from the fact that the definition of STARC metrics is symmetric in R_1 and R_2. For 3, the fact that m is a metric again implies that $d(R_1, R_3) = m(s(R_1), s(R_3)) \leq m(s(R_1), s(R_2)) + m(s(R_2), s(R_3)) = d(R_1, R_2) + d(R_2, R_3)$. This completes the proof. □

In Section 4.6, we will prove that all STARC metrics are both sound and complete. In other words, for any STARC metric d, we have that a small value of d is both necessary and sufficient for a low regret. This means that STARC metrics,

in a certain sense, exactly capture what it means for two reward functions to be similar, and that we should not expect it to be possible to significantly improve upon them. Also recall that Proposition 46 says that if two pseudometrics d_1 and d_2 are both sound and complete, then d_1 and d_2 are bilipschitz equivalent. This means that STARC metrics are unique up to bilipschitz equivalence. In particular, all STARC metrics are bilipschitz equivalent, and any other pseudometric on $\mathcal{R}$ that induces both an upper and a lower bound on worst-case regret (as we define it) must also be bilipschitz equivalent to STARC metrics.

We will next give a few concrete examples of STARC metrics. We begin by showing how to construct canonicalisation functions:

Proposition 51. *For any policy π, the function $c : \mathcal{R} \to \mathcal{R}$ given by*

$$c(R)(s, a, s') = \mathbb{E}_{S' \sim \tau(s,a)} \left[R(s, a, S') - V^\pi(s) + \gamma V^\pi(S') \right]$$

is a canonicalisation function. Here V^π is computed under the reward R given as input to c. We call this function Value-Adjusted Levelling (VAL).

Proof. To prove that c is a canonicalisation function, we must show

1. that c is linear,
2. that $c(R)$ and R differ by potential shaping and S'-redistribution, and
3. that $c(R_1) = c(R_2)$ if R_1 and R_2 differ by potential shaping and S'-redistribution.

We first show that c is linear. Given a state s, let v_s be the $|\mathcal{S}||\mathcal{A}||\mathcal{S}|$-dimensional vector where the (s', a, s'')'th dimension is given by

$$\sum_{i=0}^{\infty} \gamma^i \cdot \mathbb{P}(S_i = s', A_i = a, S_{i+1} = s''),$$

where the probability is given for a trajectory that is generated from π and τ, starting in s. Now note that $V^\pi(s) = v_s \cdot R$, where R is represented as a vector. Using these vectors $\{v_s\}$, it is possible to express c as a linear transformation.

To see that $c(R)$ and R differ by potential shaping and S'-redistribution, it is sufficient to note that V^π acts as a potential function, and that setting $R_2(s,a,s') = \mathbb{E}_{S' \sim \tau(s,a)}[R_1(s,a,S')]$ is a form of S'-redistribution.

To see that $c(R_1) = c(R_2)$ if R_1 and R_2 differ by potential shaping and S'-redistribution, first note that if R_1 and R_2 differ by potential shaping, so that $R_2(s,a,s') = R_1(s,a,s') + \gamma\Phi(s') - \Phi(s)$ for some Φ, then $V_2^\pi(s) = V_1^\pi(s) - \Phi(s)$ (Proposition 29). This means that

$$\begin{aligned} c(R_2)(s,a,s') =& \mathbb{E}[R_2(s,a,S') + \gamma \cdot V_2^\pi(S') - V_2^\pi(s)] \\ =& \mathbb{E}[R_1(s,a,S') + \gamma \cdot \Phi(S') - \Phi(s) \\ & + \gamma \cdot (V_1^\pi(S') - \Phi(S')) - (V_1^\pi(s) - \Phi(s))] \\ =& \mathbb{E}[R_1(s,a,S') + \gamma \cdot V_1^\pi(S') - V_1^\pi(s)] \\ =& c(R_1)(s,a,s'). \end{aligned}$$

It is also easy to see that $c(R_1) = c(R_2)$ if R_1 and R_2 differ by S'-redistribution. To see that $c(R_1) = c(R_2)$ only if R_1 and R_2 differ by potential shaping and S'-redistribution, first note that we have already shown that R and $c(R)$ differ by potential shaping and S'-redistribution for all R. This implies that R_1 and $c(R_1)$ differ by potential shaping and S'-redistribution, and likewise for R_2 and $c(R_2)$. Then if $c(R_1) = c(R_2)$, we can combine these transformations, and obtain that R_1 and R_2 also differ by potential shaping and S'-redistribution. □

Proposition 51 gives us an easy way to make canonicalisation functions. Moreover, while our focus in this work is on theoretical analysis, we think it is worth noting that VAL can be approximated in large-scale environments, as long as V^π can be approximated well. We next give another example of canonicalisation functions with desirable theoretical properties:

Definition 52. A canonicalisation function $c : \mathcal{R} \to \mathcal{R}$ is *minimal* for a norm n if for all R_1 we have that $n(c(R_1)) \leq n(R_2)$ for all R_2 such that R_1 and R_2 differ by potential shaping and S'-redistribution.

It is not a given that minimal canonicalisation functions exist for a given norm n, or that they are unique. For example, the minimal canonicalisation function is not unique for the L_1-norm, since its unit ball is not strictly convex. However, for any weighted L_2-norm, the minimal canonicalisation function does exist, and is unique:

Proposition 53. *For any weighted L_2-norm, a minimal canonicalisation function exists and is unique.*

Proof. Let R_0 be the reward function that is 0 for all transitions. First recall that the set of all reward functions that differ from R_0 by potential shaping and S'-redistribution forms a linear subspace of $\mathcal{R}$ (Proposition 33). Let this space be denoted by $\mathcal{Y}$, and let $\mathcal{X}$ denote the orthogonal complement of $\mathcal{Y}$ in $\mathcal{R}$. Now any reward function $R \in \mathcal{R}$ can be uniquely expressed in the form $R_\mathcal{X} + R_\mathcal{Y}$, where $R_\mathcal{X} \in \mathcal{X}$ and $R_\mathcal{Y} \in \mathcal{Y}$. Consider the function $c : \mathcal{R} \to \mathcal{R}$ where $c(R) = R_\mathcal{X}$. Now this function is a canonicalisation function such that $n(c(R)) \leq R'$ for all R' such that $c(R) = c(R')$, assuming that n is a weighted L_2-norm. To see this, we must show that

1. c is linear,
2. $c(R)$ and R differ by potential shaping and S'-redistribution for all R,
3. $c(R_1) = c(R_2)$ for all R_1 and R_2 which differ by potential shaping and S'-redistribution, and
4. $n(c(R)) \leq n(R')$ for all R' such that $c(R) = c(R')$.

It follows directly from the construction that c is linear. To see that $c(R)$ and R differ by potential shaping and S'-redistribution, simply note that $c(R) = R - R_\mathcal{Y}$, where $R_\mathcal{Y}$ is given by a combination of potential shaping and S'-redistribution of R_0. To see that $c(R_1) = c(R_2)$ if R_1 and R_2 differ by potential shaping and S'-redistribution, let $R_2 = R_1 + R'$, where R' is given by potential shaping and S'-redistribution of R_0, and let $R_1 = R_\mathcal{X} + R_\mathcal{Y}$, where $R_\mathcal{X} \in \mathcal{X}$ and $R_\mathcal{Y} \in \mathcal{Y}$. Now $c(R_1) = R_\mathcal{X}$. Moreover, $R_2 = R_\mathcal{X} + R_\mathcal{Y} + R'$. We also have that $R' \in \mathcal{Y}$, which means

that R_2 can be expressed as $R_{\mathcal{X}} + (R_{\mathcal{Y}} + R')$, where $R_{\mathcal{X}} \in \mathcal{X}$ and $(R_{\mathcal{Y}} + R') \in \mathcal{Y}$. This implies that $c(R_2) = R_{\mathcal{X}}$, so if R_1 and R_2 differ by potential shaping and S'-redistribution, then $c(R_1) = c(R_2)$. To see that $c(R_1) = c(R_2)$ only if R_1 and R_2 differ by potential shaping and S'-redistribution, first note that we have already shown that R and $c(R)$ differ by potential shaping and S'-redistribution for all R. This implies that R_1 and $c(R_1)$ differ by potential shaping and S'-redistribution, and likewise for R_2 and $c(R_2)$. Then if $c(R_1) = c(R_2)$, we can combine these transformations, and obtain that R_1 and R_2 also differ by potential shaping and S'-redistribution.

To see that $n(c(R)) \leq n(R')$ for all R' such that $c(R) = c(R')$, first note that if $c(R) = c(R')$, then $R = R_{\mathcal{X}} + R_{\mathcal{Y}}$ and $R' = R_{\mathcal{X}} + R'_{\mathcal{Y}}$, where $R_{\mathcal{X}} \in \mathcal{X}$ and $R_{\mathcal{Y}}, R'_{\mathcal{Y}} \in \mathcal{Y}$. This means that $n(c(R)) = n(R_{\mathcal{X}})$, and $n(R') = n(R_{\mathcal{X}} + R'_{\mathcal{Y}})$. Moreover, since n is a weighted L_2-norm, and since $R_{\mathcal{X}}$ and $R'_{\mathcal{Y}}$ are orthogonal, we have that $n(R_{\mathcal{X}} + R_{\mathcal{Y}}) = \sqrt{n(R_{\mathcal{X}})^2 + n(R_{\mathcal{Y}})^2} \geq n(R_{\mathcal{X}})$. This means that $n(c(R)) \leq n(R')$.

To see that this canonicalisation function is the unique minimal canonicalisation function for any weighted L_2-norm n, consider an arbitrary reward function R. Now, the set of all reward functions that differ from R by potential shaping and S'-redistribution forms an affine space of $\mathcal{R}$, and a minimal canonicalisation function must map R to a point R' in this space such that $n(R') \leq n(R'')$ for all other points R'' in that space. If n is a weighted L_2-norm, then this specifies a convex optimisation problem with a unique solution. □

Finally, we will introduce one more canonicalisation function, which will be useful for illustrative purposes:

Proposition 54. *Let $\Omega = \{\eta^\pi : \pi \in \Pi\}$ be the set of all occupancy measures, and let $c : \mathcal{R} \to \mathcal{R}$ be the function that projects each reward function onto the linear subspace of $\mathcal{R}$ that is parallel to Ω. Then c is a canonicalisation function.*

Proof. To show that c is a canonicalisation function, we must show that

1. c is linear,

2. $c(R)$ and R differ by potential shaping and S'-redistribution for all R, and

3. $c(R_1) = c(R_2)$ for all R_1 and R_2 which differ by potential shaping and S'-redistribution.

It is straightforward that c is linear, since it is a projection map. To see that R and $c(R)$ differ by potential shaping and S'-redistribution, note that there is a constant k such that $\eta^\pi \cdot R = \eta^\pi \cdot c(R) + k$ for all policies π, since $\mathrm{Im}(c)$ is parallel to Ω. This means that the policy evaluation functions of R and $c(R)$ differ by a constant k. By Proposition 29, this means that we can create a reward function R' which has the same policy evaluation function as $c(R)$, by applying potential shaping to R with a potential function such that $\Phi(s) = -k$ for all $s \in \mathrm{supp}(\mu_0)$. By Lemma 39, this implies that R' and $c(R)$ differ by potential shaping and S'-redistribution. Thus R and $c(R)$ differ by potential shaping and S'-redistribution. Finally, note that if R_1 and R_2 differ by potential shaping and S'-redistribution, then there is a constant k such that $\eta^\pi \cdot R_2 = \eta^\pi \cdot R_1 + k$ for all policies π (Proposition 29). This in turn means that $c(R_1) = c(R_2)$. □

A STARC metric can use any canonicalisation function c. Moreover, the normalisation step can use any function n that is a norm on $\mathrm{Im}(c)$. This does of course include the L_1-norm, L_2-norm, L_∞-norm, and so on. We next show that $\max_\pi J(\pi) - \min_\pi J(\pi)$ also is a norm on $\mathrm{Im}(c)$:

Proposition 55. *If c is a canonicalisation function, then the function $n : \mathcal{R} \to \mathcal{R}$ given by $n(R) = \max_\pi J(\pi) - \min_\pi J(\pi)$ is a norm on $\mathrm{Im}(c)$. Here J is computed under the reward R given as input to c.*

Proof. To show that a function n is a norm on $\mathrm{Im}(c)$, we must show that it satisfies:

1. $n(R) \geq 0$ for all $R \in \mathrm{Im}(c)$.

2. $n(R) = 0$ if and only if $R = R_0$ for all $R \in \mathrm{Im}(c)$.

3. $n(\alpha \cdot R) = \alpha \cdot n(R)$ for all $R \in \mathrm{Im}(c)$ and all scalars α.

4. $n(R_1 + R_2) \leq n(R_1) + n(R_2)$ for all $R_1, R_2 \in \text{Im}(c)$.

Here R_0 is the reward function that is 0 everywhere. It is trivial to show that Axioms 1 and 3 are satisfied by n. For Axiom 2, note that $n(R) = 0$ exactly when $\max_\pi J(\pi) = \min_\pi J(\pi)$. If R is R_0, then $J(\pi) = 0$ for all π, and so the "if" part holds straightforwardly. For the "only if" part, let R be a reward function such that $\max_\pi J(\pi) = \min_\pi J(\pi)$. Then R and R_0 induce the same policy ordering under τ and μ_0, which means that they differ by potential shaping, S'-redistribution, and positive linear scaling (Theorem 40). Moreover, since R_0 is 0 everywhere, this means that R and R_0 in fact differ by potential shaping and S'-redistribution. However, from the definition of canonicalisation functions, if $R_1, R_2 \in \text{Im}(c)$ differ by potential shaping and S'-redistribution, then it must be that $R_1 = R_2$. Hence Axiom 2 holds as well. We can show that Axiom 4 holds algebraically:

$$\begin{aligned} n(R_1 + R_2) &= \max_\pi(J_1(\pi) + J_2(\pi)) - \min_\pi(J_1(\pi) + J_2(\pi)) \\ &\leq \max_\pi J_1(\pi) + \max_\pi J_2(\pi) - \min_\pi J_1(\pi) - \min_\pi J_2(\pi) \\ &= (\max_\pi J_1(\pi) - \min_\pi J_1(\pi)) + (\max_\pi J_2(\pi) - \min_\pi J_2(\pi)) \\ &= n(R_1) + n(R_2) \end{aligned}$$

This means that $n(R) = \max_\pi J(\pi) - \min_\pi J(\pi)$ is a norm on $\text{Im}(c)$. □

For the final step we of course have that any norm is an admissible metric, though some other metrics are admissible as well. For example, if $m(R_1, R_2)$ is the *angle* between R_2 and R_2 when $R_1, R_2 \neq R_0$, and we define $m(R_0, R_0) = 0$ and $m(R, R_0) = \pi/2$ for $R \neq R_0$, then m is also admissible. To obtain a STARC metric, we then pick any canonicalisation function c, norm n, and admissible metric m, and combine them as described in Definition 49.

Some of our results will apply to *any* pseudometric on $\mathcal{R}$, and most of our other results will apply to any pseudometric on $\mathcal{R}$ that is both sound and complete (including any STARC metric). However, to obtain specific quantitative results, we will sometimes have to use a specific pseudometric. Therefore, we will use the following STARC metric as our "standard" pseudometric:

Definition 56. Let $d^{\text{STARC}}_{\tau,\gamma}$ be the STARC metric for which n is the L_2-norm, c is the canonicalisation function that is minimal for the L_2-norm, and m is the metric given by $m(x,y) = 0.5 \cdot L_2(x,y)$.

We will use $c^{\text{STARC}}_{\tau,\gamma}$ to denote the canonicalisation function of $d^{\text{STARC}}_{\tau,\gamma}$ (i.e., the minimal canonicalisation function for the L_2-norm). Moreover, we will use $s^{\text{STARC}}_{\tau,\gamma} : \mathcal{R} \to \mathcal{R}$ to denote the function that is equal to

$$\left(\frac{c^{\text{STARC}}_{\tau,\gamma}(R)}{L_2(c^{\text{STARC}}_{\tau,\gamma}(R))} \right)$$

when $L_2(c^{\text{STARC}}_{\tau,\gamma}(R)) > 0$, and $c^{\text{STARC}}_{\tau,\gamma}(R)$ otherwise.[6] We let m be equal to half the L_2-distance, to ensure that $d^{\text{STARC}}_{\tau,\gamma}$ is bounded between 0 and 1. The reason for why we will use $d^{\text{STARC}}_{\tau,\gamma}$ as our "standard" STARC metric is mainly that $d^{\text{STARC}}_{\tau,\gamma}$ is easy to work with. However, this choice is not very consequential, since all STARC metrics are bilipschitz equivalent.

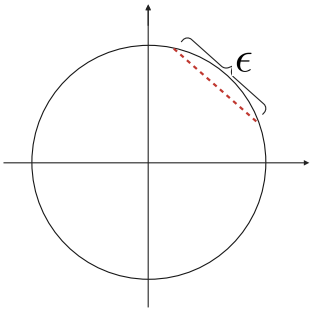
4.5 Understanding STARC Metrics

In this section, we provide a geometric intuition for how STARC metrics work. This will make it easier to understand STARC metrics, and may also make it easier to understand our proofs.

First of all, note that the space of all reward functions $\mathcal{R}$ forms an $|\mathcal{S}||\mathcal{A}||\mathcal{S}|$-dimensional vector space. Next, recall that if two reward functions R_1 and R_2 differ by (some combination of) potential shaping and S'-redistribution, then R_1 and R_2 induce the same ordering of policies. Moreover, these transformations correspond to a linear subspace of $\mathcal{R}$ (Proposition 33). A canonicalisation function is simply a linear map that removes the dimensions that are associated with potential shaping and S'-redistribution. In other words, they map $\mathcal{R}$ to an $|\mathcal{S}|(|\mathcal{A}| - 1)$-dimensional subspace of $\mathcal{R}$ in which no reward functions differ by potential shaping or S'-redistribution. The canonicalisation function that is minimal for the L_2-norm is the orthogonal map that satisfies these properties, whereas other canonicalisation

[6] This means that $d^{\text{STARC}}_{\tau,\gamma}(R_1, R_2) = 0.5 \cdot L_2(s^{\text{STARC}}_{\tau,\gamma}(R_1), s^{\text{STARC}}_{\tau,\gamma}(R_2))$.

functions are non-orthogonal. When we normalise the resulting reward functions by dividing by a norm n, we project the entire vector space onto the unit ball of n (except the zero reward, which remains at the origin). The metric m then measures the distance between the resulting reward functions on the surface of this sphere:



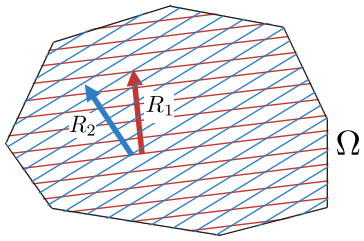
To make this more clear, it may be worth considering the case of non-sequential decision making. Suppose we have a finite set of *choices* C, and a utility function $U : C \to \mathbb{R}$. Given two distributions D_1, D_2 over C, we say that we prefer D_1 over D_2 if $\mathbb{E}_{c\sim D_1}[U(c)] > \mathbb{E}_{c\sim D_2}[U(c)]$. The set of all utility functions over C forms a $|C|$-dimensional vector space. Moreover, in this setting, it is well-known that two utility functions U_1, U_2 induce the same preferences between all possible distributions over C if and only if they differ by an affine transformation. Therefore, if we wanted to represent the set of all *non-equivalent* utility functions over C, we may consider requiring that $U(c_0) = 0$ for some $c_0 \in C$, and that $L_2(U) = 1$ unless $U(c) = 0$ for all $c \in C$. Any utility function over C is equivalent to some utility function in this set, and this set can in turn be represented as the surface of a $(|C| - 1)$-dimensional sphere, together with the origin. This is essentially analogous to the standardisation that the canonicalisation function c and the normalisation function n perform for STARC metrics. Here C is analogous to the set of all trajectories, the trajectory return function G is analogous to U, and a policy π induces a distribution over trajectories. Affine transformations of the trajectory return function, G, correspond exactly to potential shaping and positive linear scaling of R (Propositions 31 and 32). However, it is also important to note that while the cases are analogous, it

is not a direct correspondence, because not all distributions over trajectories can be realised as a policy in a given MDP.

Another perspective that may help with understanding STARC metrics comes from considering the occupancy measures of policies. Recall that the occupancy measure η^π of a policy π is the $|\mathcal{S}||\mathcal{A}||\mathcal{S}|$-dimensional vector in which the value of the (s, a, s')'th dimension is

$$\sum_{t=0}^{\infty} \gamma^t \mathbb{P}_{\xi\sim\pi}(S_t = s, A_t = a, S_{t+1} = s').$$

Also recall that $J(\pi) = \eta^\pi \cdot R$. Therefore, by computing occupancy measures, we can divide the computation of J into two parts, the first of which is independent of R, and the second of which is a linear function. Moreover, let $\Omega = \{\eta^\pi : \pi \in \Pi\}$ be the set of all occupancy measures. We now have that the policy value function J of a reward function R can be visualised as a linear function on this set. Moreover, if we have two reward functions R_1, R_2, then they can be visualised as two different linear functions on this set:



From this image, it is visually clear that the worst-case regret of maximising R_1 instead of R_2, should be proportional to the angle between the linear functions that R_1 and R_2 induce on Ω. Moreover, this is what STARC metrics measure. In particular, the function c that projects each R onto the linear subspace of $\mathcal{R}$ that is parallel to Ω is a canonicalisation function (Proposition 54). Normalising these reward functions, and measuring their distance using a metric that is bilipschitz equivalent to a norm, is bilipschitz equivalent to measuring their angle. This should in turn give an intuition for why the STARC distance between two rewards provide both an upper and lower bound on their worst-case regret.

4.6 Soundness and Completeness

In this section, we will prove that all STARC metrics are both sound and complete. We will begin by showing that they are sound. To do this, we must first prove a number of supporting lemmas:

Lemma 57. *For any rewards R_1 and R_2, and any policy π, we have that*

$$|J_1(\pi) - J_2(\pi)| \leq \left(\frac{1}{1-\gamma}\right) L_\infty(R_1, R_2).$$

Proof. This follows from straightforward algebra:

$$\begin{aligned}
&|J_1(\pi) - J_2(\pi)| \\
=&\left|\mathbb{E}_{\xi\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t R_1(S_t, A_t, S_{t+1})\right] - \mathbb{E}_{\xi\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t R_2(S_t, A_t, S_{t+1})\right]\right| \\
=&\left|\sum_{t=0}^{\infty}\gamma^t \mathbb{E}_{\xi\sim\pi}[R_1(S_t, A_t, S_{t+1}) - R_2(S_t, A_t, S_{t+1})]\right| \\
\leq&\sum_{t=0}^{\infty}\gamma^t \mathbb{E}_{\xi\sim\pi}[|R_1(S_t, A_t, S_{t+1}) - R_2(S_t, A_t, S_{t+1})|] \\
\leq&\sum_{t=0}^{\infty}\gamma^t L_\infty(R_1, R_2) = \left(\frac{1}{1-\gamma}\right) L_\infty(R_1, R_2).
\end{aligned}$$

Here the third line follows from the linearity of expectation, and the fourth line follows from Jensen's inequality. □

Thus, the L_∞-distance between two rewards bounds the difference between their policy evaluation functions. Since all norms are bilipschitz equivalent on any finite-dimensional vector space, this extends to all norms:

Lemma 58. *If p is a* norm*, then there is a positive constant K_p such that, for any reward functions R_1 and R_2, and any policy π, $|J_1(\pi) - J_2(\pi)| \leq K_p \cdot p(R_1, R_2)$.*

Proof. If p and q are norms on a finite-dimensional vector space, then there are constants k and K such that $k \cdot p(x) \leq q(x) \leq K \cdot p(x)$. Since $\mathcal{S}$ and $\mathcal{A}$ are finite, $\mathcal{R}$ is a finite-dimensional vector space. This means that there is a constant K such that $L_\infty(R_1, R_2) \leq K \cdot p(R_1, R_2)$. Together with Lemma 57, this implies that

$$|J_1(\pi) - J_2(\pi)| \leq \left(\frac{1}{1-\gamma}\right) \cdot K \cdot m(R_1, R_2).$$

Letting $K_p = \left(\frac{K}{1-\gamma}\right)$ completes the proof. □

Next, we show that if the difference between two policy evaluation functions can be bounded, then we can derive a regret bound:

Lemma 59. *Let R_1 and R_2 be reward functions, and π_1, π_2 be two policies. If $|J_1(\pi) - J_2(\pi)| \leq U$ for $\pi \in \{\pi_1, \pi_2\}$, and if $J_2(\pi_2) \geq J_2(\pi_1)$, then*

$$J_1(\pi_1) - J_1(\pi_2) \leq 2 \cdot U.$$

Proof. First note that U must be non-negative. Next, note that if $J_1(\pi_1) < J_1(\pi_2)$ then $J_1(\pi_1) - J_1(\pi_2) < 0$, and so the lemma holds. Now consider the case when $J_1(\pi_1) \geq J_1(\pi_2)$:

$$\begin{aligned} J_1(\pi_1) - J_1(\pi_2) &= J_1(\pi_1) - J_2(\pi_2) + J_2(\pi_2) - J_1(\pi_2) \\ &\leq |J_1(\pi_1) - J_2(\pi_2)| + |J_2(\pi_2) - J_1(\pi_2)| \end{aligned}$$

Our assumptions imply that $|J_2(\pi_2) - J_1(\pi_2)| \leq U$. We will next show that $|J_1(\pi_1) - J_2(\pi_2)| \leq U$ as well. Our assumptions imply that

$$\begin{aligned} &|J_1(\pi_1) - J_2(\pi_1)| \leq U \\ \implies &J_2(\pi_1) \geq J_1(\pi_1) - U \\ \implies &J_2(\pi_2) \geq J_1(\pi_1) - U \end{aligned}$$

Here the last implication uses the fact that $J_2(\pi_2) \geq J_2(\pi_1)$. A symmetric argument also shows that $J_1(\pi_1) \geq J_2(\pi_2) - U$ (recall that we assume that $J_1(\pi_1) \geq J_1(\pi_2)$). Together, this implies that $|J_1(\pi_1) - J_2(\pi_2)| \leq U$. We have thus shown that if $J_1(\pi_1) \geq J_1(\pi_2)$ then

$$|J_1(\pi_1) - J_2(\pi_2)| + |J_2(\pi_2) - J_1(\pi_2)| \leq 2 \cdot U,$$

and so the lemma holds. This completes the proof. □

Note that Lemma 57 and 59 together imply that, for any two reward functions R_1, R_2, and any two policies π_1, π_2, if $J_2(\pi_2) \geq J_2(\pi_1)$, then

$$J_1(\pi_2) - J_1(\pi_2) \leq \left(\frac{2}{1-\gamma}\right) L_\infty(R_1, R_2).$$

Moreover, Lemma 58 says that a similar bound can be derived for any norm. To turn this into a regret bound for STARC metrics, we must consider the effect of the canonicalisation function and normalisation. Our next lemma will be used to account for the difference between the size of a reward function before and after canonicalisation:

Lemma 60. *For any linear function $c : \mathbb{R}^n \to \mathbb{R}^n$ and any norm n, there is a positive constant K_n such that $n(c(v)) \leq K_n \cdot n(v)$ for all $v \in \mathbb{R}^n$.*

Proof. First consider the case when $n(v) > 0$. In this case, we can find an upper bound for $n(c(v))$ in terms of $n(v)$ by finding an upper bound for $\frac{n(c(R))}{n(R)}$. Since c is linear, and since n is absolutely homogeneous, we have that for any $v \in \mathbb{R}^n$ and any non-zero $\alpha \in \mathbb{R}$,

$$\frac{n(c(\alpha \cdot v))}{n(\alpha \cdot v)} = \left(\frac{\alpha}{\alpha}\right) \frac{n(c(v))}{n(v)} = \frac{n(c(v))}{n(v)}.$$

In other words, $\frac{n(c(v))}{n(v)}$ is unaffected by scaling of v. We may thus restrict our attention to the unit ball of n. Next, since the surface of the unit ball of n is a compact set, and since $\frac{n(c(v))}{n(v)}$ is continuous on this surface, the extreme value theorem implies that $\frac{n(c(v))}{n(v)}$ must take on some maximal value K_n on this domain. Together, the above implies that $n(c(v)) \leq K_n \cdot n(v)$ for all R such that $n(v) > 0$.

Next, suppose $n(v) = 0$. In this case, v is the zero vector. Since c is linear, this implies that $c(v) = v$, which means that $n(c(v)) = 0$ as well. Therefore, if $n(v) = 0$, then the statement holds for any K_n. In particular, it holds for the value K_n selected above. This completes the proof. □

Note that the value of K_n depends on how "tilted" $\mathrm{Im}(c)$ is. If c is an orthogonal projection (as is the case if c is the minimal canonicalisation function for an L_2-norm), then $K_n = 1$. Our next lemma has a somewhat complicated statement, but its purpose is simply to derive a bound on the difference between the policy evaluation functions J_1, J_2 of two reward functions R_1, R_2, based on the difference between the policy evaluation functions of their standardised counterparts:

Lemma 61. *Let c be a canonicalisation function, and let n be a norm on* $\mathrm{Im}(c)$*. Let R be any reward function, and let $R_S = \left(\frac{c(R)}{n(c(R))}\right)$ if $n(c(R)) > 0$, and $c(R)$ otherwise. Then $J(\pi_1) - J(\pi_2) = n(c(R)) \cdot (J_S(\pi_1) - J_S(\pi_2))$, where J_S is the policy evaluation function of R_S.*

Proof. Let us first consider the case where $n(c(R)) = 0$. Since n is a norm, $c(R)$ must be the reward function that is 0 everywhere. Since c is a canonicalisation function, we have that R and $c(R)$ have the same ordering of policies. Thus R is *trivial*, which means that $J(\pi_1) = J(\pi_2)$ for all π_1, π_2. Thus $J(\pi_1) - J(\pi_2) = 0$, and so the statement holds.

Let us next consider the case when $n(c(R)) > 0$. Let $R_C = c(R)$. Since c is a canonicalisation function, we have that R and R_C differ by potential shaping and S'-redistribution. Thus, for all π, $J_C(\pi) = J(\pi) - \mathbb{E}_{S_0 \sim \mu_0}[\Phi(S_0)]$ for some potential function Φ, where J_C is the policy evaluation function of R_C. (Proposition 29). Moreover, $J_S = J_C \cdot \left(\frac{1}{n(c(R))}\right)$. This means that

$$J_S(\pi) = \left(\frac{1}{n(c(R))}\right)(J(\pi) - \mathbb{E}_{S_0 \sim \mu_0}[\Phi(S_0)])$$

for all π. This further implies that

$$J_S(\pi_1) - J_S(\pi_2) = \left(\frac{1}{n(c(R))}\right)(J(\pi_1) - J(\pi_2))$$

since the $\mathbb{E}_{S_0 \sim \mu_0}[\Phi(S_0)]$-terms cancel out. By rearranging, we get that

$$J(\pi_1) - J(\pi_2) = n(c(R))(J_S(\pi_1) - J_S(\pi_2)).$$

This completes the proof. □

Using this, we can now finally prove that all STARC metrics are sound:

Theorem 62. *All STARC metrics are sound.*

Proof. Consider any transition function τ and any initial state distribution μ_0, and let d be a STARC metric. We wish to show that there exists a positive constant U, such that for any R_1 and R_2, and any pair of policies π_1 and π_2 such that $J_2(\pi_2) \geq J_2(\pi_1)$, we have that

$$J_1(\pi_1) - J_1(\pi_2) \leq (\max_\pi J_1(\pi) - \min_\pi J_1(\pi)) \cdot U \cdot d(R_1, R_2).$$

Recall that $d(R_1, R_2) = m(s(R_1), s(R_2))$, where m is an admissible metric. Since m is admissible, we have that $p(s(R_1), s(R_2)) \leq K_m \cdot m(s(R_1), s(R_2))$ for some norm p and constant K_m. Moreover, since p is a norm, we can apply Lemma 58 to conclude that there is a constant K_p such that for any policy π, we have that

$$|J_1^S(\pi) - J_2^S(\pi)| \leq K_p \cdot p(s(R_1), s(R_2)),$$

where J_1^S is the policy evaluation function of $s(R_1)$, and J_2^S is the policy evaluation function of $s(R_2)$. Combining this with the fact that $p(s(R_1), s(R_2)) \leq K_m \cdot m(s(R_1), s(R_2))$, we get

$$\begin{aligned} |J_1^S(\pi) - J_2^S(\pi)| &\leq K_p \cdot p(s(R_1), s(R_2)) \\ &\leq K_p \cdot K_m \cdot m(s(R_1), s(R_2)) \\ &= K_{mp} \cdot d(R_1, R_2) \end{aligned}$$

where $K_{mp} = K_p \cdot K_m$. We have thus established that, for any π, we have

$$|J_1^S(\pi) - J_2^S(\pi)| \leq K_{mp} \cdot d(R_1, R_2).$$

Let π_1 and π_2 be any two policies such that $J_2(\pi_2) \geq J_2(\pi_1)$. Note that $J_2(\pi_2) \geq J_2(\pi_1)$ if and only if $J_2^S(\pi_2) \geq J_2^S(\pi_1)$. We can therefore apply Lemma 59 and conclude that

$$J_1^S(\pi_1) - J_1^S(\pi_2) \leq 2 \cdot K_{mp} \cdot d(R_1, R_2).$$

We can now apply Lemma 61:

$$J_1(\pi_1) - J_1(\pi_2) \leq n(c(R_1)) \cdot 2 \cdot K_{mp} \cdot d(R_1, R_2).$$

We have that n is a norm on $\mathrm{Im}(c)$. Moreover, $\max_\pi J_1(\pi) - \min_\pi J_1(\pi)$ is also a norm on $\mathrm{Im}(c)$ (Proposition 55). Since $\mathrm{Im}(c)$ is a finite-dimensional vector space, this means that there is a constant K_s such that $n(c(R_1)) \leq K_s \cdot (\max_\pi J_1(\pi) - \min_\pi J_1(\pi))$ for all $R_1 \in \mathcal{R}$. Let $U = 2 \cdot K_{mp} \cdot K_s$. We have now established that, for any π_1 and π_2 such that $J_2(\pi_2) \geq J_2(\pi_1)$, we have

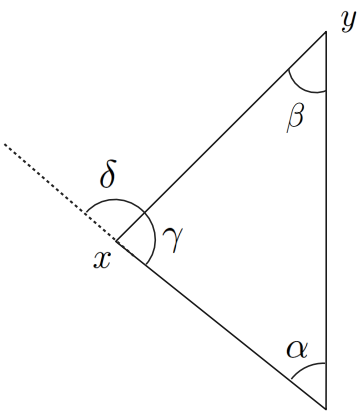
$$J_1(\pi_1) - J_1(\pi_2) \leq (\max_\pi J_1(\pi) - \min_\pi J_1(\pi)) \cdot U \cdot d(R_1, R_2).$$

This completes the proof. □

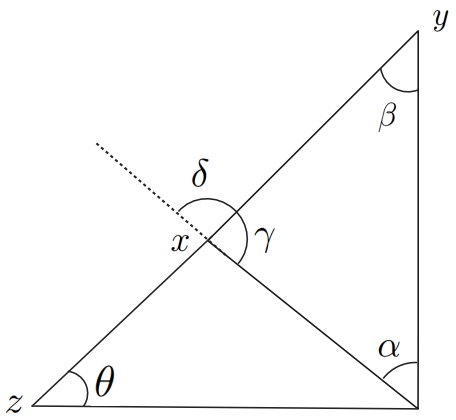
This is one of our main results for this section. Our second main result is that all STARC metrics are complete. To prove this, we must yet again first prove a number of supporting lemmas:

Lemma 63. *Let $S \subset \mathbb{R}^n$ be the boundary of a bounded convex set whose interior is non-empty and includes the origin. Then there is an $A > 0$ such that for any $x, y \in S$, if $x \neq y$ then the angle between x and $y - x$ is at least A.*

Proof. Let x, y be two arbitrary points in S such that $x \neq y$. Let α be the angle between x and y, let β be the angle between $-y$ and $x - y$, let γ be the angle between $-x$ and $y - x$, and let δ be the angle between x and $y - x$. Note that $\alpha + \beta + \gamma = \pi$, since these angles are the interior angles of the triangle whose corners lie at x, y, and the origin. We also have that $\gamma + \delta = \pi$, since these two angles add up to the angle between x and $-x$. We seek a lower bound on δ.



First note that if $\alpha > \pi/2$ then $\gamma < \pi/2$, since $\alpha + \beta + \gamma = \pi$. This means that $\delta > \pi/2$, since $\gamma + \delta = \pi$. Next, suppose $\alpha \leq \pi/2$. Since $\gamma + \delta = \pi$, we can derive a lower bound for δ by deriving an upper bound for γ. Let z be the point such that the angle between y and z is $\pi/2$, and such that x lies on the line segment between z and y. Let θ be the angle between $-z$ and $y - z$.



Now elementary trigonometry tells us that $\gamma < \pi/2 + \theta$.[7] By deriving an upper bound for θ, we thus obtain an upper bound for γ (and hence a lower bound for δ).

Note that $\theta = \arctan(L_2(y)/L_2(z))$. Moreover, since S is the boundary of some set X, and since the interior of X is non-empty, there must be some $\ell > 0$ such that $L_2(x) \geq \ell$ for all $x \in S$. Moreover, since X is bounded, there must be some $u \geq \ell$ such that $L_2(x) \leq u$ for all $x \in S$. We have that $L_2(y) \leq u$, since $y \in S$.

It may be that $z \notin S$. However, since S is the boundary of a *convex* set, it must still be the case that $L_2(z) \geq \ell$. To see this, suppose $L_2(z) < \ell$, and let z' be the point in S such that $z' = a \cdot z$ for some $a \in \mathbb{R}^+$. $L_2(z') \geq \ell$, since $s' \in S$, and so $L_2(z') > L_2(z)$. Consider the triangle that lies between z', y, and the origin. Since S is the boundary of a convex set X, every point that lies in the interior of this triangle must lie in the interior of X. But if $L_2(z') > L_2(z)$, then x lies in the

[7] In particular, $\gamma = \pi - \alpha - \beta$, and $\beta = \pi/2 - \theta$. Thus γ is maximised when $\alpha = 0$, in which case $\gamma = \pi - (\pi/2 - \theta) = \pi/2 + \theta$. Moreover, if $\alpha = 0$ then $x = y$, which by assumption is not the case. Hence $\gamma < \pi/2 + \theta$.

interior of this triangle. This is a contradiction, since x lies on the boundary of X. Thus $L_2(z) \geq \ell$.

We thus have that $\theta \leq \arctan(u/\ell)$, which means that $\gamma < \pi/2 + \arctan(u/\ell)$, and thus that $\delta > \pi - (\pi/2 + \arctan(u/\ell)) = \pi/2 - \arctan(u/\ell)$. Since this value does not depend on x or y, we have that the angle δ between x and $y - x$ is at least $\pi/2 - \arctan(u/\ell)$ for all $x, y \in S$ such that $x \neq y$, and such that the angle α between x and y is less than or equal to $\pi/2$. Also recall that if $\alpha > \pi/2$ then $\delta > \pi/2$. Since $u/\ell > 0$, we have that $\pi/2 - \arctan(u/\ell) < \pi/2$, and so $\delta > \pi/2 - \arctan(u/\ell)$ for all $x, y \in S$ such that $x \neq y$. Finally, since $\arctan(x) < \pi/2$, we have that $\pi/2 - \arctan(u/\ell) > 0$. Setting $A = \pi/2 - \arctan(u/\ell)$ thus completes the proof. □

Using this, we can now show that we can get a lower bound on the *angle* between two standardised reward functions in terms of their STARC-distance:

Lemma 64. *For any STARC metric d, there exist an $\ell_1 \in \mathbb{R}^+$ such that the angle θ between $s(R_1)$ and $s(R_2)$ satisfies $\ell_1 \cdot d(R_1, R_2) \leq \theta$ for all R_1, R_2 for which neither $s(R_1)$ or $s(R_2)$ is R_0.*

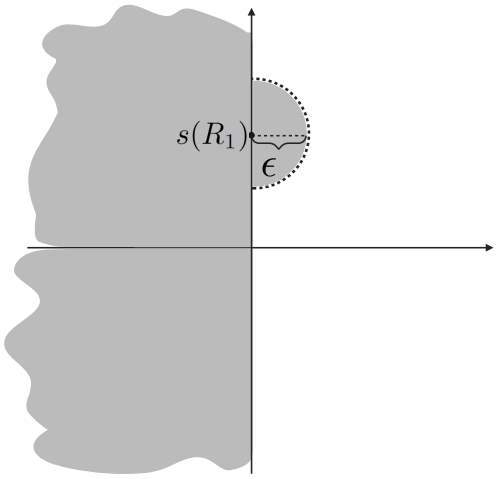
Proof. Let d be an arbitrary STARC-metric, and let R_1 and R_2 be two arbitrary reward functions for which neither $s(R_1)$ or $s(R_2)$ is R_0. Recall that $d(R_1, R_2) = m(s(R_1), s(R_2))$, where m is a metric that is bilipschitz equivalent to some norm. Since all norms are bilipschitz equivalent on any finite-dimensional vector space, this means that m is bilipschitz equivalent to the L_2-norm. Thus, there are positive constants p, q such that

$$p \cdot m(s(R_1), s(R_2)) \leq L_2(s(R_1), s(R_2)) \leq q \cdot m(s(R_1), s(R_2)).$$

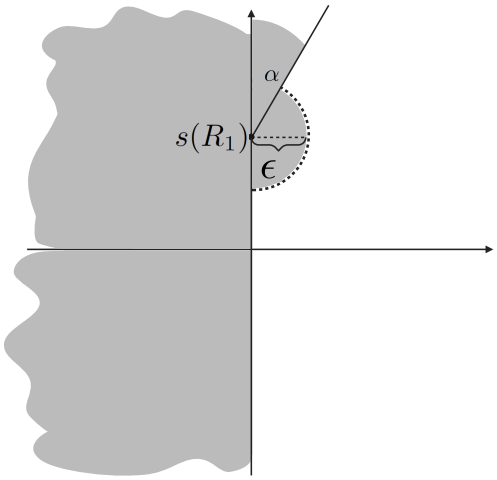
In particular, the L_2-distance between $s(R_1)$ and $s(R_2)$ is at least $\epsilon = p \cdot d(R_1, R_2)$. For the rest of our proof, it will be convenient to assume that $\epsilon < L_2(s(R_1))$; this can be ensured by picking a p that is sufficiently small.[8]

Let us plot the plane which contains $s(R_1)$, $s(R_2)$, and the origin, and orient it so that $s(R_1)$ points straight up, and so that $s(R_2)$ is not on the left-hand side:

[8] Specifically, we need to pick a p that is no bigger than $\max_R L_2(s(R)) / \max_{R_1, R_2} d(R_1, R_2)$.

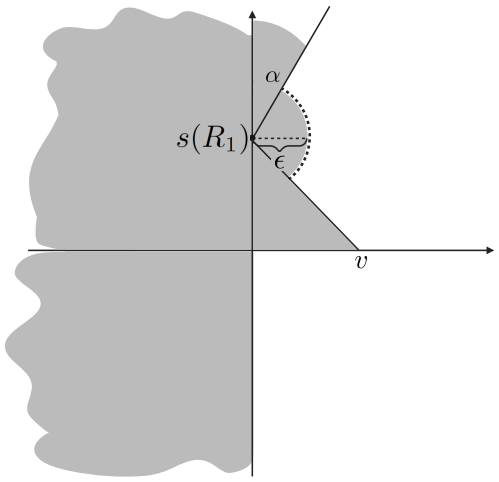


Since the distance between $s(R_1)$ and $s(R_2)$ is at least ϵ, and since $s(R_2)$ is not on the left-hand side, we know that $s(R_2)$ cannot be inside of the region shaded grey in the figure above (though it may be on the boundary). Moreover, as per Lemma 63, we know that there is an $\alpha > 0$ (named A in the statement of Lemma 63) such that the angle between $s(R_1)$ and $s(R_2) - s(R_1)$ is at least α. This means that we also can rule out the following region:



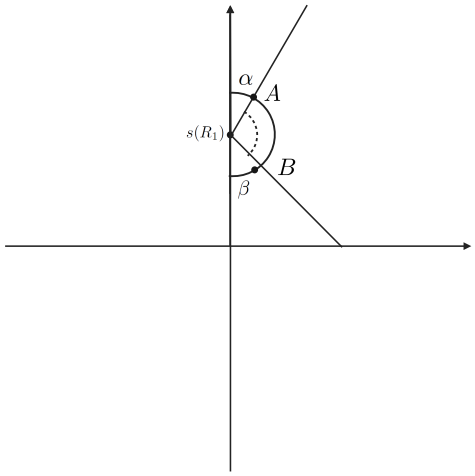
Let v be the element of $\mathrm{Im}(s)$ that is perpendicular to $s(R_1)$ and lies on the right-hand side in the figure.[9] Since $\mathrm{Im}(s)$ is the boundary of a convex set, we know that $s(R_2)$ cannot lie within the triangle that lies between $s(R_1)$, v, and the origin:

[9]That is, v is perpendicular to $s(R_1)$, lies on a plane with $s(R_1)$, $s(R_2)$, and the origin, and minimises the angle to $s(R_2)$ among the two vectors that satisfy the previous constraints.



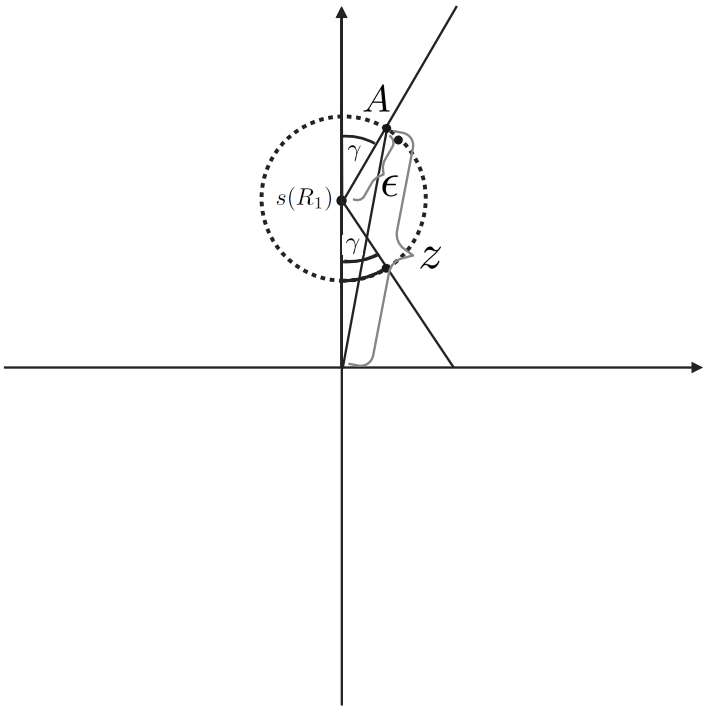
Since $\text{Im}(s)$ is compact, we know that there is a vector a in $\text{Im}(s)$ whose L_2-norm is bigger than all other vectors in $\text{Im}(s)$, and a (non-zero) vector b in $\text{Im}(s)$ whose L_2-norm is smaller than all other (non-zero) vectors in $\text{Im}(s)$, by the extreme value theorem. From this, we can infer that the angle between $-s(R_1)$ and $v - s(R_1)$ is at least $\beta = \arctan(|b|/|a|)$. Also note that $\beta > 0$.

We now have everything we need to derive a lower bound on the angle θ between $s(R_1)$ and $s(R_2)$. First note that θ can be no smaller than the angle between $s(R_1)$ and the points marked A and B in the figure below (whichever is smaller):



To make things easier, replace α and β with $\gamma = \min(\alpha, \beta)$. Since this makes the shaded region smaller, we still have that $s(R_2)$ cannot be in the interior of the

new shaded region. Moreover, in this case, we know that the angle between $s(R_1)$ and $s(R_2)$ is no smaller than the angle θ' between $s(R_1)$ and the point marked A:



Deriving this angle is now just a matter of trigonometry. Let z denote $L_2(A)$. Using the law of sines, we have that:

$$\frac{\epsilon}{\sin(\theta')} = \frac{z}{\sin(\pi - \gamma)} = \frac{z}{\sin(\gamma)}$$

From this, we get that

$$\begin{aligned} \theta' &= \arcsin\left(\left(\frac{\epsilon}{z}\right)\sin(\gamma)\right) \\ &\geq \left(\frac{\epsilon}{z}\right)\sin(\gamma) \end{aligned}$$

Moreover, it is also straightforward to find an upper bound z' for z. Specifically,

$$\begin{aligned} z^2 &= L_2(s(R_1))^2 + \epsilon^2 - 2L_2(s(R_1))\epsilon\cos(\pi - \gamma) \\ &= L_2(s(R_1))^2 + \epsilon^2 + 2L_2(s(R_1))\epsilon\cos(\gamma). \end{aligned}$$

Since $\epsilon < L_2(s(R_1))$, and since $\gamma < \pi/2$, this means that

$$\begin{aligned} z &< \sqrt{2L_2(s(R_1))^2 + 2L_2(s(R_1))^2\cos(\gamma)} \\ &= L_2(s(R_1))\sqrt{2(1+\cos(\gamma))} \end{aligned}$$

Moreover, since $\mathrm{Im}(s)$ is compact, there is a vector $a \in \mathrm{Im}(s)$ whose L_2-norm is bigger than all other vectors in $\mathrm{Im}(s)$. We thus know that

$$z < z' = L_2(a)\sqrt{2(1+\cos(\gamma))}.$$

Putting this together, we have that

$$\theta \geq \theta' \geq \left(\frac{\epsilon}{z'}\right)\sin(\gamma) = m(s(R_1), s(R_2)) \cdot p \cdot \left(\frac{\sin(\gamma)}{z'}\right).$$

Setting $\ell_1 = p \cdot \left(\frac{\sin(\gamma)}{z'}\right)$ thus completes the proof, since the values of p, γ, and z' do not depend on R_1 or R_2. □

Finally, before we can give the full proof, we will also need the following:

Lemma 65. *For any invertible matrix $M : \mathbb{R}^n \to \mathbb{R}^n$ there is an $\ell_2 \in (0, 1]$ such that for any $v, w \in \mathbb{R}^n$, the angle θ' between Mv and Mw satisfies $\theta' \geq \ell_2 \cdot \theta$, where θ is the angle between v and w.*

Proof. We will first prove that this holds in the 2-dimensional case, and then extend this proof to the general n-dimensional case.

Let M be an arbitrary invertible matrix $\mathbb{R}^2 \to \mathbb{R}^2$. First note that we can factor M via Singular Value Decomposition into three matrices U, Σ, V, such that $M = U\Sigma V^\top$, where U and V are orthogonal matrices, and Σ is a diagonal matrix with non-negative real numbers on the diagonal. Since M is invertible, we also have that Σ cannot have any zeroes along its diagonal. Next, recall that orthogonal matrices preserve angles. This means that we can restrict our focus to just Σ.[10]

[10] If there are vectors x, y such that the angle between x and y is θ and the angle between Mx and My is θ', then there are vectors v, w such that the angle between x and y is θ and the angle between Σv and Σw is θ', and vice versa.

Let α and β be the singular values of M. We may assume, without loss of generality, that

$$\Sigma = \begin{pmatrix} \alpha & 0 \\ 0 & \beta \end{pmatrix}.$$

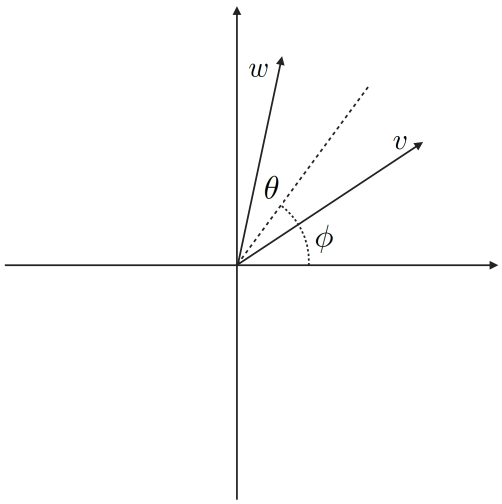
Moreover, since scaling the x and y-axes uniformly will not affect the angle between any vectors after multiplication, we can instead equivalently consider the matrix

$$\Sigma = \begin{pmatrix} \alpha/\beta & 0 \\ 0 & 1 \end{pmatrix}.$$

Let $v, w \in \mathbb{R}^2$ be two arbitrary vectors with angle θ, and let θ' be the angle between Σv and Σw. We will derive a lower bound on θ' expressed in terms of θ. Moreover, since the angle between v and w is not affected by their magnitude, we will assume (without loss of generality) that both v and w have length 1 (under the L_2-norm).

First, note that if $\theta = \pi$ then $v = -w$. This means that $\Sigma v = -\Sigma w$, since Σ is a linear transformation, which in turn means that $\theta' = \pi$. Thus $\theta' \geq \ell_2 \cdot \theta$ as long as $\ell_2 \leq 1$. Next, assume that $\theta < \pi$.

We may assume (without loss of generality) that the angle between v and the x-axis is no bigger than the angle between w and the x-axis. Let ϕ be the angle between the x-axis and the vector that is in the middle between v and w. This means that we can express v as $(\cos(\phi - \theta/2), \sin(\phi - \theta/2))$ and w as $(\cos(\phi + \theta/2), \sin(\phi + \theta/2))$. Moreover, since reflection along either of the axes will not change the angle between either v and w or Σv and Σw, we may assume (without loss of generality) that $\phi \in [0, \pi/2]$. For convenience, we will also let $\sigma = \alpha/\beta$.



(Note that we can visualise the action of Σ as scaling the x-axis in the figure above by σ.)

We now have that $\Sigma v = (\sigma \cos(\phi - \theta/2), \sin(\phi - \theta/2))$ and $\Sigma w = (\sigma \cos(\phi + \theta/2), \sin(\phi + \theta/2))$. Using the dot product, we get that $\cos(\theta')$ equals

$$\frac{\sigma^2 \cos(\phi - \theta/2)\cos(\phi + \theta/2) + \sin(\phi - \theta/2)\sin(\phi + \theta/2)}{\sqrt{\sigma^2 \cos^2(\phi - \theta/2) + \sin^2(\phi - \theta/2)}\sqrt{\sigma^2 \cos^2(\phi + \theta/2) + \sin^2(\phi + \theta/2)}}.$$

We next note that if $\theta \in [0, \pi)$ and $\phi \in [0, \pi/2]$, then the derivative of this expression with respect to ϕ can only be 0 when $\phi \in \{0, \pi/2\}$.[11] This means that $\cos(\theta')$ must be maximised or minimised when ϕ is either 0 or $\pi/2$, which in turn means that the angle θ' must be minimised or maximised when ϕ is either 0 or $\pi/2$.

It is now easy to see that if $\sigma > 1$ then θ' is minimised when $\phi = 0$, and that if $\sigma < 1$ then θ' is minimised when $\phi = \pi/2$. Moreover, if $\phi = \pi/2$, then

$$\theta' = 2 \arctan\left(\frac{\sigma \cos(\pi/2 - \theta/2)}{\sin(\pi/2 - \theta/2)}\right) = 2 \arctan\left(\sigma \tan(\theta/2)\right),$$

which in turn is greater than $\theta \cdot \sigma$ when $\sigma < 1$.[12] Similarly, if $\phi = 0$, then

$$\theta' = 2 \arctan\left(\frac{\sin(\theta/2)}{\sigma \cos(\theta/2)}\right) = 2 \arctan\left(\sigma^{-1} \tan(\theta/2)\right),$$

[11] For example, this may be verified using tools such as Wolfram Alpha.

[12] To see this, let $x = \tan(\theta/2)$. Now $2 \arctan(\sigma \tan(\theta/2)) > \sigma \cdot \theta$ for all $\theta \in [0, \pi)$ if and only if $\arctan(\sigma x) > \sigma \cdot \arctan(x)$ for all $x \geq 0$. This is true, since arctan is strictly concave on $[0, \infty)$.

which is in turn greater than $\sigma^{-1} \cdot \theta$ when $\sigma > 1$. In either case, we thus have

$$\theta' \geq \theta \cdot \min(\sigma, \sigma^{-1}) = \theta \cdot \min(\beta/\alpha, \alpha/\beta).$$

Finally, if $\sigma = 1$, then of course $\theta' = \theta$ and $\beta/\alpha = \alpha/\beta = 1$, and so $\theta' \geq \theta \cdot \min(\beta/\alpha, \alpha/\beta)$ in this case as well. We have therefore show that, for any invertible matrix $M : \mathbb{R}^2 \to \mathbb{R}^2$, there exists a positive constant $\min(\beta/\alpha, \alpha/\beta)$, where α and β are the singular values of M, such that if $v, w \in \mathbb{R}^2$ have angle θ, then the angle between Mv and Mw is at least $\theta \cdot \min(\beta/\alpha, \alpha/\beta)$.

To generalise this to the general n-dimensional case, let $v, w \in \mathbb{R}^n$ be two arbitrary vectors. Consider the 2-dimensional linear subspace given by $S = \text{span}(v, w)$, and note that $M(S)$ also is a 2-dimensional linear subspace of $\mathbb{R}^n$ (since M is linear and invertible). The linear transformation which M induces between S and $M(S)$ is isomorphic to a linear transformation $M' : \mathbb{R}^2 \to \mathbb{R}^2$.[13] We can thus apply our previous result for the two-dimensional case, and conclude that if the angle between v and w is θ, then the angle between Mv and Mw is at least $\theta \cdot \min(\beta/\alpha, \alpha/\beta)$, where α and β are the singular values of M'. Next, note that the singular values of M' cannot be smaller than the smallest singular values of M or bigger than the biggest singular values of M. We can therefore let $\ell_2 = \alpha/\beta$, where α is the smallest singular value of M and β is the greatest singular value of M, and conclude that the angle between Mv and Mw must be at least $\ell_2 \cdot \theta$. Since the value of ℓ_2 does not depend on v or w, this completes the proof. □

Using this, we can now prove that all STARC metrics are complete:

[13] To see this, let A be an orthonormal matrix that rotates $\mathbb{R}^2$ to align with S, and let B be an orthonormal matrix that rotates $M(S)$ to align with $\mathbb{R}^2$. Now $M' = BMA$ is an invertible linear transformation $\mathbb{R}^2 \to \mathbb{R}^2$. Moreover, since orthonormal matrices preserve the angles between vectors, we have that $v, w \in S$ have angle θ and $Mv, Mw \in M(S)$ have angle θ', if and only if $A^{-1}v, A^{-1}w \in \mathbb{R}^2$ have angle θ and $BMv, BMw \in \mathbb{R}^2$ have angle θ'. Note that $M'A^{-1}v = BMv$ and $M'A^{-1}w = BMw$. This means that there are $v, w \in S$ such that v, w have angle θ and Mv, Mw have angle θ', if and only if there are $v', w' \in \mathbb{R}^2$ such that v', w' have angle θ and $M'v'$ and $M'w'$ have angle θ' (with $v' = A^{-1}v$ and $w' = A^{-1}w$).

Theorem 66. *All STARC metrics are complete.*

Proof. Let d be an arbitrary STARC metric. We need to show that there exists a positive constant L such that, for any reward functions R_1 and R_2, there are two policies π_1, π_2 with $J_2(\pi_2) \geq J_2(\pi_1)$ and

$$J_1(\pi_1) - J_1(\pi_2) \geq L \cdot (\max_{\pi} J_1(\pi) - \min_{\pi} J_1(\pi)) \cdot d(R_1, R_2),$$

and if both R_1 and R_2 are trivial, then we have that $d(R_1, R_2) = 0$.

Let c be the canonicalisation function of d, and let $s : \mathcal{R} \to \mathcal{R}$ be the function such that $s(R) = c(R)/n(c(R))$ if $n(c(R)) \neq 0$, and $c(R)$ otherwise, where n is the norm used in the normalisation step of c.

First note that the last condition holds straightforwardly. If both R_1 and R_2 are trivial, then $c(R_1) = c(R_2) = R_0$, which implies that $d(R_1, R_2) = 0$.

For the first condition, let us first consider the case when R_1 is trivial (and R_2 may be trivial or non-trivial). In this case the left-hand side is 0 for all π_1 and π_2. Moreover, $\max_\pi J_1(\pi) - \min_\pi J_1(\pi) = 0$, and so the right-hand side is also 0 (for any value of L). In this case, the inequality is therefore satisfied for any L.

Let us next consider the case where R_2 is trivial, but where R_1 is not. In this case, $J_2(\pi_2) \geq J_2(\pi_1)$ for all π_1 and π_2, which means that $\max_{\pi_1,\pi_2:J_2(\pi_2)\geq J_2(\pi_1)} J_1(\pi_1) - J_1(\pi_2) = \max_\pi J_1(\pi) - \min_\pi J_1(\pi)$. Therefore, the inequality is satisfied as long as we pick an L such that $L \cdot d(R_1, R_2) \leq 1$ for all R_1 and all trivial R_2. In other words, we need that $L \leq 1/\max_R d(R, R_0)$. This can be ensured by picking an L that is sufficiently small (noting that any STARC metric d is bounded).

Finally, let us consider the case where neither R_1 or R_2 is trivial, i.e., the case where $\max_\pi J_1(\pi) - \min_\pi J_1(\pi) > 0$ and $\max_\pi J_2(\pi) - \min_\pi J_2(\pi) > 0$. Let $m : \Pi \to \mathbb{R}^{|\mathcal{S}||\mathcal{A}||\mathcal{S}|}$ be the function that takes a policy π and returns its occupancy measure η^π, and let $\Omega = \text{Im}(m)$. Note that m implicitly depends on τ and γ. We will use d to derive a lower bound on the angle θ between the level sets of J_1 and J_2 in Ω. We will then show that Ω contains an open set with a certain diameter. From this, we can find two policies that incur a certain amount of regret.

First, by Lemma 64, there exists an ℓ_1 such that for any non-trivial R_1 and R_2, the angle between $s(R_1)$ and $s(R_2)$ is at least $\ell_1 \cdot d(R_1, R_2)$. To make our proof easier, we will assume that we pick an ℓ_1 that is small enough to ensure that $\ell_1 \cdot d(R_1, R_2) \leq \pi/2$ for all R_1, R_2. Since d is bounded, this is possible.

Note that $s(R_1)$ and $s(R_2)$ may not be parallel with Ω, which means that the angle between $s(R_1)$ and $s(R_2)$ may not be the same as the angle between the level sets of $s(R_1)$ and $s(R_2)$ in Ω. Therefore, consider the matrix M that projects $\mathcal{R}$ onto the linear subspace of $\mathcal{R}$ that is parallel to Ω, where c is the canonicalisation function of d. Now the angle between $Ms(R_1)$ and $Ms(R_2)$ is the same as the angle between the level sets of the linear functions which J_1 and J_2 induce on Ω. Moreover, recall that M is a canonicalisation function (Proposition 54). This means that the elements of $\mathrm{Im}(M)$ and $\mathrm{Im}(c)$ can be put in a one-to-one correspondence, and so M is invertible when viewed as a function from $\mathrm{Im}(c)$. Also note that $s(R_1), s(R_2) \in \mathrm{Im}(c)$. We can therefore apply Lemma 64, and conclude that there exists an $\ell_2 \in (0, 1]$, such that the angle θ between the normal vectors (and hence the level sets) of $s(R_1)$ and $s(R_2)$ in Ω is at least $\ell_2 \cdot \ell_1 \cdot d(R_1, R_2)$. Moreover, since $\ell_1 \cdot d(R_1, R_2) \leq \pi/2$, and since $\ell_2 \leq 1$, we have that $\ell_2 \cdot \ell_1 \cdot d(R_1, R_2) \leq \pi/2$.

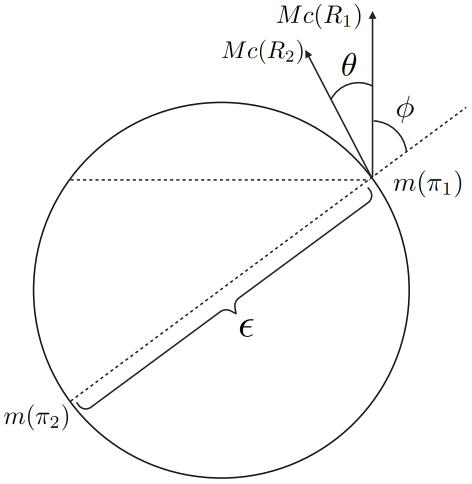
This gives us that, for any two policies π_1, π_2, we have:

$$\begin{aligned} J_1(\pi_1) - J_1(\pi_2) &= J_1^C(\pi_1) - J_1^C(\pi_2) \\ &= c(R_1) \cdot m(\pi_1) - c(R_1) \cdot m(\pi_2) \\ &= c(R_1) \cdot (m(\pi_1) - m(\pi_2)) \\ &= M(c(R_1)) \cdot (m(\pi_1) - m(\pi_2)) \\ &= L_2(M(c(R_1))) \cdot L_2(m(\pi_1) - m(\pi_2)) \cdot \cos(\phi) \end{aligned}$$

where ϕ is the angle between $M(c(R_1))$ and $m(\pi_1) - m(\pi_2)$, and J_1^C is the evaluation function of $c(R_1)$. Note that the first and fourth line follow from Proposition 29. We can thus derive a lower bound on worst-case regret by deriving a lower bound for the greatest value of this expression.

By Lemma 38, we have that Ω contains a set that is open in the smallest affine space which contains Ω. This means that there is an ϵ, such that Ω contains a

sphere of diameter ϵ. We will show that we always can find two policies within this sphere that incur a certain amount of regret. Consider the 2-dimensional cut which goes through the middle of this sphere and is parallel with the normal vectors of the level sets of J_1 and J_2 in Ω. The intersection between this cut and the ϵ-sphere forms a 2-dimensional circle with diameter ϵ. Let π_1, π_2 be the two policies for which $m(\pi_1)$ and $m(\pi_2)$ lie opposite to each other on this circle, and satisfy that $J_2(\pi_1) = J_2(\pi_2)$ (or, equivalently, that $Mc(R_1) \cdot m(\pi_1) = Mc(R_1) \cdot m(\pi_2)$). Without loss of generality, we may assume that $J_1(\pi_1) \geq J_1(\pi_2)$.



Now note that $L_2(m(\pi_1) - m(\pi_2)) = \epsilon$. Moreover, recall that the angle θ between $Mc(R_1)$ and $Mc(R_2)$ is at least $\theta' = \ell_1 \cdot \ell_2 \cdot d(R_1, R_2)$, and that this quantity is at most $\pi/2$. This means that the angle ϕ is at most $\pi/2 - \theta'$, and so $\cos(\phi)$ is at least $\cos(\pi/2 - \theta') = \cos(\pi/2 - \ell_1 \cdot \ell_2 \cdot d(R_1, R_2))$. This means that we have two policies π_1, π_2 where $J_2(\pi_2) = J_2(\pi_1)$ and such that

$$\begin{aligned} J_1(\pi_1) - J_1(\pi_2) &= L_2(M(c(R_1))) \cdot L_2(m(\pi_1) - m(\pi_2)) \cdot \cos(\phi) \\ &\geq L_2(M(c(R_1))) \cdot \epsilon \cdot \cos(\pi/2 - \ell_2 \cdot \ell_1 \cdot d(R_1, R_2)) \\ &= L_2(M(c(R_1))) \cdot \epsilon \cdot \sin(\ell_2 \cdot \ell_1 \cdot d(R_1, R_2)). \end{aligned}$$

Note that $\sin(x) \geq x \cdot 2/\pi$ when $x \leq \pi/2$, and that $\ell_2 \cdot \ell_1 \cdot d(R_1, R_2) \leq \pi/2$. Putting this together, we have that there exists π_1, π_2 with $J_2(\pi_2) = J_2(\pi_1)$ such that

$$J_1(\pi_1) - J_1(\pi_2) \geq L_2(M(c(R_1))) \cdot \left(\frac{\epsilon \cdot \ell_1 \cdot \ell_2 \cdot 2}{\pi} \right) \cdot d(R_1, R_2).$$

Next, note that, if p is a norm and M is an invertible matrix, then $p \circ M$ is also a norm. Furthermore, recall that $\max_\pi J_1(\pi) - \min_\pi J_1(\pi)$ is a norm on $\mathrm{Im}(c)$, when c is a canonicalisation function (Proposition 55). Since all norms are equivalent on a finite-dimensional vector space, this means that there must exist a positive constant ℓ_3 such that

$$L_2(M(c(R_1))) \geq \ell_3 \cdot (\max_\pi J_1^C(\pi) - \min_\pi J_1^C(\pi)) = \ell_3 \cdot (\max_\pi J_1(\pi) - \min_\pi J_1(\pi)).$$

We can therefore set $L \leq (\epsilon \cdot \ell_1 \cdot \ell_2 \cdot \ell_3 \cdot 2/\pi)$, and obtain the result we want:

$$J_1(\pi_1) - J_1(\pi_2) \geq L \cdot (\max_\pi J_1(\pi) - \min_\pi J_1(\pi)) \cdot d(R_1, R_2).$$

This completes the proof. □

We have thus proven that all STARC metrics are complete, which is our second main result for this section. Since STARC metrics are both sound and complete, they provide a canonical method for quantifying the differences between reward functions.

I conclude that other human beings have feelings like me, because, first, they have bodies like me, which I know, in my own case, to be the antecedent condition of feelings; and because, secondly, they exhibit the acts, and other outward signs, which in my own case I know by experience to be caused by feelings.

— John Stuart Mill, 1865.

5 Partial Identifiability

In this chapter, we present our results about the partial identifiability of the reward function in IRL, relative to the standard behavioural models. Specifically, we derive the ambiguity of the Boltzmann-rational model, the MCE model, and the optimality model. We also discuss the ambiguity tolerance of a number of different applications, and analyse the question of transfer learning. Note that our results in this section cover both the case where reward functions are compared using equivalence relations, and the case where they are compared using pseudometrics.

5.1 Invariances of Intermediate Objects

Many types of policies can be computed via some intermediate objects. For example, the Boltzmann-rational policy can be computed by first computing the optimal advantage function $A^\star$, and then applying a softmax function. Moreover, recall that for any two reward objects $f : \mathcal{R} \to X$ and $g : \mathcal{R} \to Y$, if there exists a function $h : X \to Y$ such that $h \circ f = g$, then $\text{Am}(f) \preceq \text{Am}(g)$ (Lemma 4). This means that if $g(R)$ can be computed by first computing some intermediate object $f(R)$, then g inherits all of the invariances of f. For example, $b_{\tau,\mu_0,\beta}$ inherits the invariances of $A^\star$. For this reason, it will be useful to catalogue the invariances of a number of such objects, which we will do in this section.

We begin by deriving the invariances of different forms of Q-functions:

Lemma 67. *For any transition function τ, any discount factor γ, and any policy π, the Q-function Q^π determines R up to $S'\mathrm{R}_\tau$.*

Proof. Recall that Q^π is the only function which satisfies the Bellman equation (Equation 2.2) for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$:

$$Q^\pi(s,a) = \mathbb{E}_{S' \sim \tau(s,a), A' \sim \pi(S')} \left[R(s,a,S') + \gamma \cdot Q^\pi(S', A') \right].$$

This equation can be rewritten as

$$\mathbb{E}_{S' \sim \tau(s,a)} \left[R(s,a,S') \right] = Q^\pi(s,a) - \gamma \cdot \mathbb{E}_{S' \sim \tau(s,a), A' \sim \pi(S')} \left[Q^\pi(S', A') \right].$$

Since Q^π is the only function which satisfies this equation for all $s \in \mathcal{S}, a \in \mathcal{A}$, we have that the values of the left-hand side for each $s \in \mathcal{S}, a \in \mathcal{A}$ together determine Q^π, and vice versa. Since the left-hand side values are preserved by S'-redistribution of R, and no other transformations, we have that Q^π is preserved by S'-redistribution of R, and no other transformations. □

Lemma 68. *For any transition function τ and any discount factor γ, the optimal Q-function $Q^\star$ determines R up to $S'\mathrm{R}_\tau$.*

Proof. Analogous to Lemma 67, noting that $Q^\star$ is the only function which satisfies the Bellman optimality equation for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$:

$$Q^\star(s,a) = \mathbb{E}_{S' \sim \tau(s,a), A' \sim \pi(S')} \left[R(s,a,S') + \gamma \cdot \max_{a' \in \mathcal{A}} Q^\pi(S', a') \right].$$

□

Lemma 69. *For any transition function τ, any discount γ, and any weight α, the soft Q-function Q^{S}_α determines R up to $S'\mathrm{R}_\tau$.*

Proof. Analogous to Lemma 67, noting that Q^{S}_α is the only function which satisfies the following modified Bellman equation for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$:

$$Q^{\mathrm{S}}_\alpha(s,a) = \mathbb{E}_{S' \sim \tau(s,a)} \left[R(s,a,S') + \gamma \alpha \log \sum_{a' \in \mathcal{A}} \exp\left(\frac{1}{\alpha}\right) Q^{\mathrm{S}}_\alpha(S', a') \right].$$

□

By Lemma 4, this means that any type of policy which can be derived from one of these three Q-functions will be invariant to S'-redistribution. We next derive the invariances of different forms of advantage functions:

Lemma 70. *For any transition function τ, any discount γ, and any policy π, the advantage function A^π determines R up to $\mathrm{PS}_\gamma \bigodot S'\mathrm{R}_\tau$.*

Proof. First, recall that A^π can be derived from Q^π, via $A^\pi(s,a) = Q^\pi(s,a) - \mathbb{E}_{A\sim\pi(s)}[Q(s,A)]$. Since Q^π is invariant to S'-redistribution (Lemma 67), this means that A^π is invariant to S'-redistribution (Lemma 4). Next, recall that if R_1 and R_2 differ by potential shaping with Φ, then $Q_1(s,a) = Q_2(s,a) - \Phi(s)$ (Proposition 29). This means that

$$\begin{aligned} A_1^\pi(s,a) &= Q_1^\pi(s,a) - \mathbb{E}_{A\sim\pi(s)}[Q_1(s,A)] \\ &= Q_2^\pi(s,a) - \Phi(s) - \mathbb{E}_{A\sim\pi(s)}[Q_2(s,A) - \Phi(s)] \\ &= Q_2^\pi(s,a) - \mathbb{E}_{A\sim\pi(s)}[Q_2(s,A)] \\ &= A_2^\pi(s,a). \end{aligned}$$

Together, this means that A^π is invariant to both S'-redistribution and potential shaping.

We next need to show that A^π is not invariant to any transformations that cannot be expressed as a combination of S'-redistribution and potential shaping. Let R_1 and R_2 be two reward functions such that $A_1^\pi = A_2^\pi$, and let Q_1^π, Q_2^π be their Q-functions. Define

$$\Phi(s) := \mathbb{E}_{A\sim\pi(s)}[Q_1^\pi(s,A) - Q_2^\pi(s,A)]$$

and let

$$R_3(s,a,s') = R_2(s,a,s') + \gamma\cdot\Phi(s) - \Phi(s').$$

Now R_2 and R_3 differ by potential shaping (with Φ). Moreover, by Proposition 29, we have that $Q_3^\pi = Q_1^\pi$. Therefore, by Lemma 67, we have that R_3 and R_1 differ by S'-redistribution. This implies that R_1 and R_2 differ by a combination of potential shaping and S'-redistribution. □

Lemma 71. *For any transition function τ and any discount γ, the optimal advantage function $A^\star$ determines R up to $\mathrm{PS}_\gamma \bigodot S'\mathrm{R}_\tau$.*

Proof. Analogous to Lemma 70. □

By Lemma 4, this means that any type of policy which can be derived from one of these two advantage functions will be invariant to S'-redistribution and potential shaping.

5.2 Invariances of Policies

In this section, we derive the invariances of the three behavioural models that are most common in the current IRL literature. We first note that the softmax function is invariant to constant shift, and to no other transformations:

Proposition 72. *Let $v, w \in \mathbb{R}^n$ be two vectors, and let $\beta \in \mathbb{R}^+$. Then*

$$\frac{\exp \beta v_i}{\sum_{j=1}^n \exp \beta v_j} = \frac{\exp \beta w_i}{\sum_{j=1}^n \exp \beta w_j}$$

for all $i \in \{1 \dots n\}$ if and only if there is a constant scalar c such that $v_i = w_i + c$ for all $i \in \{1 \dots n\}$.

Proof. For the first direction, suppose there is a constant scalar c such that $v_i = w_i + c$ for all $i \in \{1 \dots n\}$. Then

$$\begin{aligned}\frac{\exp \beta v_i}{\sum_{j=1}^n \exp \beta v_j} &= \frac{\exp \beta(w_i + c)}{\sum_{j=1}^n \exp \beta(w_j + c)} \\ &= \frac{\exp(\beta c) \cdot \exp \beta w_i}{\exp(\beta c) \cdot \sum_{j=1}^n \exp \beta w_j} \\ &= \frac{\exp \beta w_i}{\sum_{j=1}^n \exp \beta w_j}.\end{aligned}$$

For the other direction, suppose

$$\frac{\exp \beta v_i}{\sum_{j=1}^n \exp \beta v_j} = \frac{\exp \beta w_i}{\sum_{j=1}^n \exp \beta w_j}$$

for all $i \in \{1 \dots n\}$. Note that this can be rewritten as follows:

$$\begin{aligned} \frac{\exp \beta v_i}{\exp \beta w_i} &= \frac{\sum_{j=1}^{n} \exp \beta v_j}{\sum_{j=1}^{n} \exp \beta w_j} \\ \exp(\beta v_i - \beta w_i) &= \frac{\sum_{j=1}^{n} \exp \beta v_j}{\sum_{j=1}^{n} \exp \beta w_j} \\ v_i - w_i &= \left(\frac{1}{\beta}\right) \log \left(\frac{\sum_{j=1}^{n} \exp \beta v_j}{\sum_{j=1}^{n} \exp \beta w_j}\right) \end{aligned}$$

Since the right-hand side of this expression does not depend on i, it follows that $v_i - w_i$ is constant for all i. □

We can now derive the invariances of the Boltzmann-rational model:

Theorem 73. *For any transition function τ, discount γ, and temperature β, we have that $b_{\tau,\gamma,\beta}$ determines R up to* $\mathrm{PS}_\gamma \bigodot S'\mathrm{R}_\tau$.

Proof. Let $a_{\tau,\gamma} : \mathcal{R} \to (\mathcal{S} \times \mathcal{A} \to \mathbb{R})$ be the function that takes R and returns the optimal advantage function $A^\star$ for R, given transition function τ and discount factor γ. We will show that $\mathrm{Am}(b_{\tau,\gamma,\beta}) = \mathrm{Am}(a_{\tau,\gamma})$, by showing that the Boltzmann-rational policy can be derived from the optimal advantage function, and vice versa.

First, recall that the Boltzmann-rational policy is given by applying the softmax function (with temperature β) to the optimal Q-function, $Q^\star$, in each state. Moreover, since the the softmax function is invariant to constant shift (Proposition 72), and since $A^\star$ and $Q^\star$ differ by constant shift in each state (given by $V^\star$), we have that the Boltzmann-rational policy also can be obtained by applying the softmax function with temperature β to the optimal advantage function in each state. This means that there exists a function h such that $b_{\tau,\gamma,\beta} = h \circ a_{\tau,\gamma}$. Thus, by Lemma 4, we have that $\mathrm{Am}(a_{\tau,\gamma}) \preceq \mathrm{Am}(b_{\tau,\gamma,\beta})$.

For the other direction, recall that any softmax function is invariant to constant shift, and no other transformations (Proposition 72). Since the Boltzmann-rational policy can be derived from $A^\star$ by applying a softmax function in each state, this means that if $b_{\tau,\gamma,\beta}(R_1) = b_{\tau,\gamma,\beta}(R_2)$, then there is a function $B : \mathcal{S} \to \mathbb{R}$ such that $A_1^\star(s,a) = A_2^\star(s,a) + B(s)$ for all s, a. Moreover, since $\max_{a \in \mathcal{A}} A^\star(s,a) = 0$

for all s, it follows that $B(s) = 0$ for all s, which means that $A_1^\star = A_2^\star$. In other words, if $b_{\tau,\gamma,\beta}(R_1) = b_{\tau,\gamma,\beta}(R_2)$ then $a_{\tau,\gamma}(R_1) = a_{\tau,\gamma}(R_2)$, which means that there exists a function h such that $a_{\tau,\gamma} = h \circ b_{\tau,\gamma,\beta}$. Thus, by Lemma 4, we have that $\mathrm{Am}(b_{\tau,\gamma,\beta}) \preceq \mathrm{Am}(a_{\tau,\gamma})$.

Since $\mathrm{Am}(a_{\tau,\gamma}) \preceq \mathrm{Am}(b_{\tau,\gamma,\beta})$ and $\mathrm{Am}(b_{\tau,\gamma,\beta}) \preceq \mathrm{Am}(a_{\tau,\gamma})$, we have that $\mathrm{Am}(b_{\tau,\gamma,\beta}) = \mathrm{Am}(a_{\tau,\gamma})$. In other words, since the Boltzmann-rational policy can be derived from the optimal advantage function, and vice versa, it must be the case that they have the same invariances. Applying Lemma 71 completes the proof. □

Stated differently, $\mathrm{Am}(b_{\tau,\gamma,\beta})$ is given by $\mathrm{PS}_\gamma \bigodot S'\mathrm{R}_\tau$, so two reward functions have the same Boltzmann-rational policy if and only if they differ by potential shaping and S'-redistribution. We next consider the MCE model:

Theorem 74. *For any transition function τ, discount γ, and weight α, we have that $c_{\tau,\gamma,\alpha}$ determines R up to $\mathrm{PS}_\gamma \bigodot S'\mathrm{R}_\tau$.*

Proof. First, recall that the MCE policy is given by applying the softmax function with temperature $(1/\alpha)$ to the soft Q-function Q^{S}_α in each state. Next, recall that any softmax function is invariant to constant shift, and no other transformations (Proposition 72). This means that $c_{\tau,\gamma,\alpha}$ is invariant to all transformations that induce a constant shift of Q^{S}_α in each state, and no other transformations.

The first direction is straightforward; Lemma 69 and Proposition 30 together imply that S'-redistribution and potential shaping results in a constant shift of Q^{S}_α in all states. This means that $c_{\tau,\gamma,\alpha}$ is invariant to these transformations.

For the other direction, let R_1 and R_2 be two reward functions such that the corresponding soft Q-functions satisfy $Q^{\mathrm{S}}_{\alpha,1}(s,a) = Q^{\mathrm{S}}_{\alpha,2}(s,a) + B(s)$ for some function $B : \mathcal{S} \to \mathbb{R}$. Recall that $Q^{\mathrm{S}}_{\alpha,1}$ is the unique function which satisfies the following equation for all s and a:

$$Q^{\mathrm{S}}_{\alpha,1}(s,a) = \mathbb{E}_{S' \sim \tau(s,a)} \left[R(s,a,S') + \gamma\alpha \log \sum_{a' \in \mathcal{A}} \exp\left(\frac{1}{\alpha}\right) Q^{\mathrm{S}}_{\alpha,1}(S',a') \right].$$

This can be rewritten as

$$\mathbb{E}\left[R_1(s,a,S')\right] = Q^{\mathrm{S}}_{\alpha,1}(s,a) - \mathbb{E}\left[\gamma\alpha \log \sum_{a' \in A} \exp\left(\frac{1}{\alpha}\right) Q^{\mathrm{S}}_{\alpha,1}(S',a') \right].$$

We can now rewrite the right-hand side as follows:

$$\begin{aligned}
&Q^{\mathrm{S}}_{\alpha,1}(s,a) - \mathbb{E}\left[\gamma\alpha \log \sum_{a' \in A} \exp\left(\frac{1}{\alpha}\right) Q^{\mathrm{S}}_{\alpha,1}(S',a')\right] \\
=&Q^{\mathrm{S}}_{\alpha,2}(s,a) + B(s) - \mathbb{E}\left[\gamma\alpha \log \sum_{a' \in A} \exp\left(\frac{1}{\alpha}\right) \left(Q^{\mathrm{S}}_{\alpha,2}(S',a') + B(S')\right)\right] \\
=&Q^{\mathrm{S}}_{\alpha,2}(s,a) + B(s) - \mathbb{E}\left[\gamma\alpha \log \left(\sum_{a' \in A} \exp\left(\frac{1}{\alpha}\right) Q^{\mathrm{S}}_{\alpha,2}(S',a')\right) + \gamma B(S')\right] \\
=&\mathbb{E}\left[R_2(s,a,S') + B(s) - \gamma B(S')\right] .
\end{aligned}$$

Now set $\Phi(s) = -B(s)$, and we can see that the difference between R_1 and R_2 is described by potential shaping and S'-redistribution. □

Stated differently, $\mathrm{Am}(c_{\tau,\gamma,\alpha})$ is given by $\mathrm{PS}_\gamma \odot S'\mathrm{R}_\tau$, so two reward functions have the same MCE policy if and only if they differ by potential shaping and S'-redistribution. We next consider the optimality model:

Theorem 75. *For any transition function τ and discount γ, we have that $o^\star_{\tau,\gamma}$ determines R up to $\mathrm{OP}_{\tau,\gamma}$.*

Proof. Immediate from Theorem 34, since $o^\star_{\tau,\gamma}(R_1) = o^\star_{\tau,\gamma}(R_2)$ if and only if R_1 and R_2 have the same optimal policies. □

Stated differently, $\mathrm{Am}(o^\star_{\tau,\gamma})$ is given by $\mathrm{OP}_{\tau,\gamma}$, so two reward functions have the same maximally supportive optimal policies if and only if they differ by an optimality-preserving transformation. These results exactly characterise the partial identifiability of the reward function R under IRL which uses any of these three behavioural models.

5.3 Ambiguity Tolerance and Applications

Now that we have derived the ambiguity of R under each of the three standard behavioural models, it may be worth reflecting on the implications of these results. First of all, both $b_{\tau,\gamma,\beta}$ and $c_{\tau,\gamma,\alpha}$ determine R up to S'-redistribution (with τ) and potential shaping (with γ). From this, we can straightforwardly derive the following:

Corollary 76. *If f is $b_{\tau,\gamma,\beta}$ or $c_{\tau,\gamma,\alpha}$, then we have that:*

1. $\mathrm{Am}(f) \preceq \mathrm{ORD}_{\tau,\gamma}$.
2. $\mathrm{Am}(f) \preceq \mathrm{OPT}_{\tau,\gamma}$.
3. *If $d^{\mathcal{R}}$ is a pseudometric on $\mathcal{R}$ that is both sound and complete, then the upper and lower diameter of $\mathrm{Am}(f)$ under $d^{\mathcal{R}}$ is 0.*

Proof. As per Theorem 73 and 74, if f is either $b_{\tau,\gamma,\beta}$ or $c_{\tau,\gamma,\alpha}$, and $f(R_1) = f(R_2)$, then R_1 and R_2 differ by a transformation in $\mathrm{PS}_\gamma \bigodot S'\mathrm{R}_\tau$. As per theorem 40, this implies that $R_1 \equiv_{\mathrm{ORD}_{\tau,\gamma}} R_2$, and so $\mathrm{Am}(f) \preceq \mathrm{ORD}_{\tau,\gamma}$. Since $\mathrm{ORD}_{\tau,\gamma} \preceq \mathrm{OPT}_{\tau,\gamma}$, we also have that $\mathrm{Am}(f) \preceq \mathrm{OPT}_{\tau,\gamma}$. Finally, as per Proposition 45, all sound and complete pseudometrics metrics have the property that $d^{\mathcal{R}}(R_1, R_2) = 0$ if $R_1 \equiv_{\mathrm{ORD}_{\tau,\gamma}} R_2$. Thus the upper (and hence also the lower) diameter of $\mathrm{Am}(f)$ under $d^{\mathcal{R}}$ is 0. □

This means that for any transition function τ, any discount factor γ, and any true reward function $R^\star$, if an IRL algorithm $\mathcal{L}$ for Boltzmann-rational policies or MCE policies is trained on data that in fact is generated by a Boltzmann-rational policy or an MCE policy, then $\mathcal{L}$ will converge to a reward function R_H such that $R^\star$ and R_H have the same policy ordering (and optimal policies) under τ and γ, and that for any STARC metric $d^{\mathcal{R}}$, we have that $d^{\mathcal{R}}(R^\star, R_H) = 0$. This is good; it means that the ambiguity of these models is unproblematic.

Of course, there are a few caveats here that it is important to be cognisant of. First of all, this result relies on the assumption that the training data in fact comes from a Boltzmann-rational policy or MCE policy (i.e., that there is no misspecification). In reality, this assumption is unrealistic. In Chapters 6 and 7, we will loosen this assumption. Moreover, we are only guaranteed that $R^\star$ and R_H have the same policy order under τ and γ. In other words, we assume that R_H will be applied in the same environment where it is learnt (or, stated differently, that there is no distributional shift after the learning process). In Section 5.4, we will loosen

this assumption, and see that we fail to obtain similar guarantees in that setting. Nonetheless, even with these caveats, Theorems 73 and 74 are still good news.

Our results also show that the invariances of $o^\star_{\tau,\gamma}$ preserve $\mathrm{OPT}_{\tau,\gamma}$. This is perhaps obvious — the information that is contained in an optimal policy is of course sufficient to construct an optimal policy. Nonetheless, it is good to assimilate this result into our framework, and express it in the same terminology as our other results. Moreover, this result is of course also subject to the caveat that the training data in fact must come from an optimal policy, and the caveat that it only applies for the τ and γ that were used during training. Next, we will show that the invariances of $o^\star_{\tau,\gamma}$ do *not* preserve $\mathrm{ORD}_{\tau,\gamma}$, except in highly constrained environments:

Proposition 77. *If $o_{\tau,\gamma} \in \mathcal{O}_{\tau,\gamma}$, then unless $|\mathcal{S}| = 1$ and $|\mathcal{A}| = 2$, there are reward functions R_1, R_2 such that $o^\star_{\tau,\gamma}(R_1) = o^\star_{\tau,\gamma}(R_2)$ but $R_1 \not\equiv_{\mathrm{ORD}_{\tau,\gamma}} R_2$.*

Proof. If $|\mathcal{S}| \geq 2$ or $|\mathcal{A}| \geq 3$, then there exists uncountably many reward functions that do not have the same ordering of policies (this is immediate from Theorem 40). Moreover, $\mathrm{Im}(o_{\tau,\gamma})$ is finite. By the pigeonhole principle, this means that there must exist reward functions R_1, R_2 such that $o_{\tau,\gamma}(R_1) = o_{\tau,\gamma}(R_2)$ but $R_1 \not\equiv_{\mathrm{ORD}_{\tau,\gamma}} R_2$. □

We can thus summarise our results about the ambiguity of $o^\star_{\tau,\gamma}$ as follows:

Corollary 78. *Unless $|\mathcal{S}| = 1$ and $|\mathcal{A}| = 2$, we have that:*

1. $\mathrm{Am}(o^\star_{\tau,\gamma}) \not\preceq \mathrm{ORD}_{\tau,\gamma}$.
2. $\mathrm{Am}(o^\star_{\tau,\gamma}) \preceq \mathrm{OPT}_{\tau,\gamma}$.
3. *If $d^{\mathcal{R}}$ is a pseudometric on $\mathcal{R}$ that is both sound and complete, then the lower diameter of $\mathrm{Am}(o^\star_{\tau,\gamma})$ under $d^{\mathcal{R}}$ is 0, but the upper diameter is greater than 0.*

Proof. The first part follows from Proposition 77, and the second part follows from Theorem 75. For the third part, first note that Proposition 45 implies that if $d^{\mathcal{R}}$ is both sound and complete, then $d^{\mathcal{R}}(R_1, R_2) = 0$ if and only if $R_1 \equiv_{\mathrm{ORD}_{\tau,\gamma}} R_2$. Thus the fact that $\mathrm{Am}(o^\star_{\tau,\gamma}) \not\preceq \mathrm{ORD}_{\tau,\gamma}$ implies that the upper diameter of $\mathrm{Am}(o^\star_{\tau,\gamma})$ under $d^{\mathcal{R}}$ is greater than 0. To see that the lower diameter is 0, consider the

reward function R_0 that is 0 everywhere. Then $o^\star_{\tau,\gamma}(R_0)$ must indicate that all actions are optimal in all states, which means any reward function R such that $o^\star_{\tau,\gamma}(R) = o^\star_{\tau,\gamma}(R_0)$ must be trivial. All trivial reward functions have the same policy order, and so $d^{\mathcal{R}}(R, R_0) = 0$. □

Note that there are some special cases where $o^\star_{\tau,\gamma}(R)$ does allow us to infer the policy order of R, even if $|\mathcal{S}| \geq 2$ or $|\mathcal{A}| \geq 3$. As a simple example, if we have a one-state MDP with three actions a_1, a_2, a_3, and $o^\star_{\tau,\gamma}(R)$ shows that action a_1 and a_2 are optimal, then we can also infer the policy order of R. Alternatively, if $o^\star_{\tau,\gamma}(R)$ shows that all actions are optimal in all states, then all policies must have the same value — this is why the lower diameter of $\mathrm{Am}(o^\star_{\tau,\gamma})$ is 0. Nonetheless, these cases are marginal, and in most situations, we will not be able to infer the policy order of R from $o^\star_{\tau,\gamma}(R)$. Also note that the exact value of the upper diameter will depend on which pseudometric we use, as well as on the transition function τ and discount factor γ. Calculating this value exactly would be quite difficult, but we expect it to typically be quite large (since two reward functions may have the same optimal policies, and yet have wildly different policy orderings).

Before moving on, let us also briefly note that the behavioural models in $\mathcal{O}_{\tau,\gamma}$ other than $o^\star_{\tau,\gamma}$ in fact are too ambiguous to identify even the correct equivalence class of $\mathrm{OPT}_{\tau,\gamma}$:

Theorem 79. *If $o \in \mathcal{O}_{\tau,\gamma}$ but $o \neq o^\star_{\tau,\gamma}$, then $\mathrm{Am}(o_{\tau,\gamma}) \not\preceq \mathrm{OPT}_{\tau,\gamma}$.*

Proof. This can be demonstrated by a pigeonhole argument. Specifically, the codomain of each $o \in \mathcal{O}_{\tau,\gamma}$ has $(2^{|\mathcal{A}|} - 1)^{|\mathcal{S}|}$ elements, and there are $(2^{|\mathcal{A}|} - 1)^{|\mathcal{S}|}$ $\mathrm{OPT}_{\tau,\gamma}$-equivalence classes. This means that if $\mathrm{Am}(o) \preceq \mathrm{OPT}_{\tau,\gamma}$, then there must be a one-to-one correspondence between $\mathrm{OPT}_{\tau,\gamma}$-equivalence classes and elements of o's codomain, so that there for each equivalence class $C \in \mathrm{OPT}_{\tau,\gamma}$ is a $y_C \in \mathrm{Im}(o)$ such that $o(R) = y_C$ if and only if $R \in C$. Further, say that if $f, g : X \to \mathcal{P}(Y)$ are set-valued functions, then $f \subseteq g$ if $f(x) \subseteq g(x)$ for all $x \in X$, and $f \subset g$ if $f \subseteq g$ but $g \not\subseteq f$. Then if $o \in \mathcal{O}_{\tau,\gamma}$ we have that $o(R) \subseteq o^\star_{\tau,\gamma}(R)$ for all R — a policy is optimal if and only if it takes only optimal actions, but it need not take all optimal actions.

Moreover, if $o \neq o^\star_{\tau,\gamma}$ then there is an R_1 such that $o(R_1) \subset o^\star_{\tau,\gamma}(R_1)$. Let R_2 be a reward function so that $o^\star_{\tau,\gamma}(R_2) = o(R_1)$ — for any function $\mathcal{S} \to \mathcal{P}(\mathcal{A}) - \varnothing$, there is a reward function for which those are the optimal actions, so there is always some R_2 such that $o^\star_{\tau,\gamma}(R_2) = o(R_1)$. Now either $o(R_2) = o(R_1)$ or $o(R_2) \subset o(R_1)$, since all actions that are optimal under R_2 are optimal under R_1. In the first case, since $o(R_1) = o(R_2)$ but $R_1 \not\equiv_{\mathrm{OPT}_{\tau,\gamma}} R_2$, we have that $\mathrm{Am}(o) \not\preceq \mathrm{OPT}_{\tau,\gamma}$. In the second case, let R_3 be a reward function so that $o^\star_{\tau,\gamma}(R_3) = o(R_2)$, and repeat the same argument. Since there can only be a finite sequence $o(R_n) \subset \cdots \subset o(R_2) \subset o(R_1)$, we have that we must eventually find two R_n, R_{n-1} such that $o(R_n) = o(R_{n-1})$ but $R_n \not\equiv_{\mathrm{OPT}_{\tau,\gamma}} R_{n-1}$. This means that it cannot be the case that $\mathrm{Am}(o) \preceq \mathrm{OPT}_{\tau,\gamma}$. □

As described in Section 3.1, we can use the invariances of different reward objects to place them in a lattice structure, which graphically explains the relationship between their respective ambiguity and ambiguity tolerance — see Figure 5.1. Other data sources can be placed in the same graph, using similar techniques to what we have used in this section.

5.4 Transfer Learning

It is interesting to consider the setting where a reward function is learnt in one MDP, but used in a different MDP. For example, we may learn the reward under one transition function τ_1, but wish to use it under another transition function τ_2. Alternatively, the observed agent may discount using one discount factor γ_1, but we wish to use the reward with a different discount factor γ_2. In this section, we will demonstrate that it is impossible to guarantee robust transfer in this setting. We first derive the following lemma:

Lemma 80. *Let $r : \mathcal{S}\times\mathcal{A} \to \mathbb{R}$ be any function, R_1 any reward function, and τ_1, τ_2 any transition functions. Then there exists a reward function R_2 such that $\mathbb{E}_{S'\sim\tau_1(s,a)}[R_2(s,a,S')] = \mathbb{E}_{S'\sim\tau_1(s,a)}[R_1(s,a,S')]$ for all s, a, and such that $\mathbb{E}_{S'\sim\tau_2(s,a)}[R_2(s,a,S')] = r(s,a)$ for all s, a such that $\tau_1(s,a) \neq \tau_2(s,a)$.*

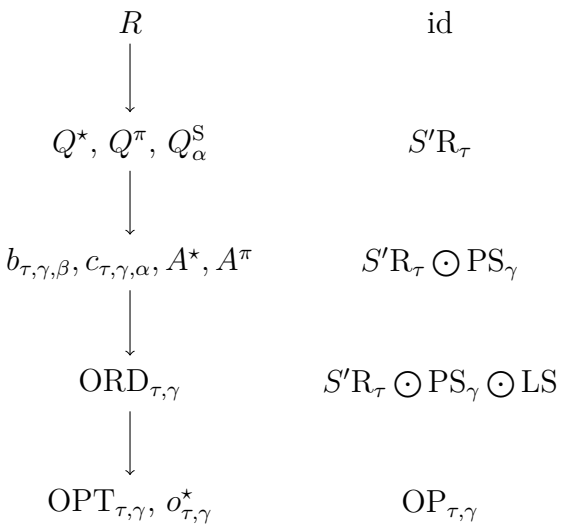


Figure 5.1: This figure summarises our results from Chapter 5. On the left-hand side, we list several reward objects and equivalence relations on $\mathcal{R}$. We write $f \to g$ if $\mathrm{Am}(f) \preceq \mathrm{Am}(g)$. Since ambiguity refinement is transitive and antisymmetric, this lets us place all reward objects in a lattice structure. Using this structure, we can read out several important relationships graphically: if $f \to g$, then a data source that is based on g is at least as ambiguous as a data source based on f, the information contained in a data source based on f is sufficient to derive the value of g as an application, and it is in principle possible to compute g based on f. Note that the lattice structure in this case forms a linear order — this is a special property of the reward objects and equivalence relations we have studied, and does not hold in general. On the right-hand side of the figure we list the reward transformations that characterise the ambiguity of the reward objects to the left.

Proof. The requirement that R_2 is produced from R_1 by S'-redistribution under τ is satisfied if, for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$,

$$\mathbb{E}_{S' \sim \tau(s,a)} \left[R_1(s, a, S') \right] = \mathbb{E}_{S' \sim \tau(s,a)} \left[R_2(s, a, S') \right] .$$

Let $s \in \mathcal{S}$ and $a \in \mathcal{A}$ be any state and action such that $\tau_1(s,a) \neq \tau_2(s,a)$. Let $\vec{\tau_1}_{s,a}$ and $\vec{\tau_2}_{s,a}$ be $\tau_1(s,a)$ and $\tau_2(s,a)$ expressed as vectors, and let $\vec{R_1}_{s,a}$ be the vector where $\vec{R_1}^{(i)}_{s,a} = R_1(s,a,s_i)$. The question is then if there is an analogous vector $\vec{R_2}_{s,a}$ such that:

$$\vec{\tau_1}_{s,a} \cdot \vec{R_2}_{s,a} = \vec{\tau_1}_{s,a} \cdot \vec{R_1}_{s,a} ,$$
$$\vec{\tau_2}_{s,a} \cdot \vec{R_2}_{s,a} = r(s,a).$$

Since $\vec{\tau_1}_{s,a}$ and $\vec{\tau_2}_{s,a}$ differ and are valid probability distributions, they are linearly independent (recall also that $|\mathcal{A}| \geq 2$). Therefore, the system of equations always

has a solution for $\vec{R}_{2_{s,a}}$. Form the required R_2 as R_1 modified to have the values of $\vec{R}_{2_{s,a}}$ in these states where the τ_1 and τ_2 differ. □

To unpack this, let R_1 be the true reward function, τ_1 be the transition dynamics of the training environment, and τ_2 be the transition dynamics of the deployment environment. Lemma 80 then says, roughly, that if the training data is invariant to S'-redistribution, and τ_1 and τ_2 differ for enough states, then the learnt reward function is essentially unconstrained in the deployment environment. Specifically, for every reward function R_1 there exists a reward function R_2 such that R_1 and R_2 are indistinguishable in the training environment, but such that R_2 may have any value for any state-action pair for which $\tau_1 \neq \tau_2$. This means that no guarantees can be obtained. Moreover, note that Lemmas 67-69 and Lemma 4 imply that this result extends to *any* object that can be computed from a Q-function, which is a very broad class. Lemma 80 then suggests that any such data source is too ambiguous to guarantee transfer to a different environment. From this, we can immediately derive the following:

Theorem 81. *If f_{τ_1} is invariant to S'-redistribution with τ_1, and $\tau_1 \neq \tau_2$, then we have that* $\mathrm{Am}(f_{\tau_1}) \not\trianglelefteq \mathrm{OPT}_{\tau_2,\gamma}$.

Proof. Let R_1 be an arbitrary reward. If $\tau_1 \neq \tau_2$, then there exists some s, a such that $\tau_1(s,a) \neq \tau_2(s,a)$. Using the construction in Lemma 80, we can find a reward function R_2 such that R_1 and R_2 differ by S'-redistribution with τ_1, and such that $A^\star_2(s,a)$ has any arbitrary value when computed under τ_2 (and any discount γ). In particular, if $a \notin \mathrm{argmax}_{a'} A^\star_1(s,a')$ under τ_2 and γ, then we can let $a \in \mathrm{argmax}_{a'} A^\star_2(s,a')$ under τ_2 and γ, and vice versa. This means that $\mathrm{argmax}_a A^\star_1(s,a) \neq \mathrm{argmax}_a A^\star_2(s,a)$ under τ_2 and γ, and so $R_1 \not\equiv_{\mathrm{OPT}_{\tau_2,\gamma}} R_2$. However, R_1 and R_2 differ by S'-redistribution with τ_1, and so $f_{\tau_1(R_1)} = f_{\tau_1(R_2)}$. □

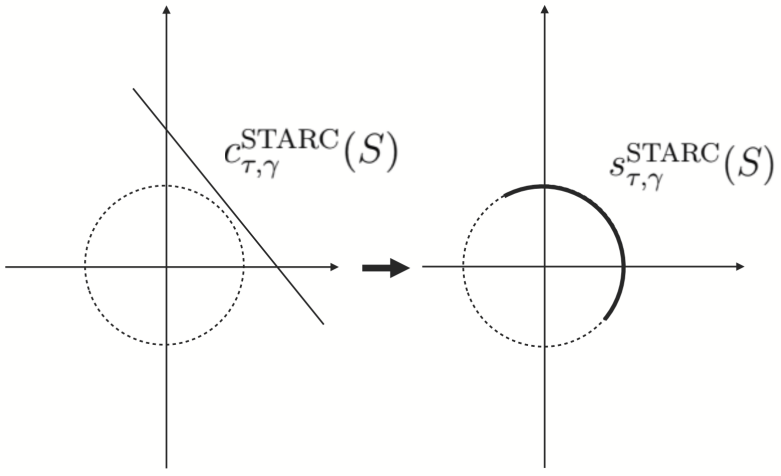
We can also extend this result to a stronger statement, expressed in terms of Definition 6. To do this, we will need the following lemma:

Lemma 82. *Let f be a reward object such that, for every reward R there exists a reward $R^\dagger$ such that $R^\dagger$ is non-trivial, and such that $f(R) = f(R + \alpha R^\dagger)$ for all $\alpha \in \mathbb{R}$. Then the lower and upper diameter of* $\mathrm{Am}(f)$ *under* $d^{\mathrm{STARC}}_{\tau,\gamma}$ *is 1.*

Proof. Let R be an arbitrary reward function, and let S be the set given by

$$S = \{R + \alpha R^\dagger : \alpha \in \mathbb{R}\}.$$

Note that S forms a line through $\mathcal{R}$, and let us consider what happens when the canonicalisation and normalisation of $d^{\mathrm{STARC}}_{\tau,\gamma}$ is applied to S. Specifically, recall that $c^{\mathrm{STARC}}_{\tau,\gamma}$ only collapses dimensions along which every reward differs by potential shaping and S'-redistribution. Moreover, R and $R + R'$ differ by potential shaping and S'-redistribution if and only if R' is trivial. Since $R^\dagger$ is non-trivial, this means that $c^{\mathrm{STARC}}_{\tau,\gamma}(S)$ forms a line through $\mathrm{Im}(c^{\mathrm{STARC}}_{\tau,\gamma})$. After the normalisation step, we have that $c^{\mathrm{STARC}}_{\tau,\gamma}(S)$ is projected onto the unit ball of n, where n is the norm used in the normalisation step of $d^{\mathrm{STARC}}_{\tau,\gamma}$. If $c^{\mathrm{STARC}}_{\tau,\gamma}(S)$ intersects the origin, then $s^{\mathrm{STARC}}_{\tau,\gamma}(S)$ will contain two points that are on the opposite sides of $\mathrm{Im}(s^{\mathrm{STARC}}_{\tau,\gamma})$, and these points have an L_2-distance of 2. If $c^{\mathrm{STARC}}_{\tau,\gamma}(S)$ does not intersect the origin, then $s^{\mathrm{STARC}}_{\tau,\gamma}(S)$ forms an arc along the surface of $\mathrm{Im}(s^{\mathrm{STARC}}_{\tau,\gamma})$. For every $\epsilon > 0$, there are two points on the far ends of this arch whose L_2-distance is at least $2 - \epsilon$. Recall that the STARC-distance between R_1 and R_2 is *half* of the L_2-distance between $s^{\mathrm{STARC}}_{\tau,\gamma}(R_1)$ and $s^{\mathrm{STARC}}_{\tau,\gamma}(R_2)$.



Since R was chosen arbitrarily, we have that there for any $x \in \text{Im}(f)$ and $\epsilon > 0$ exists reward functions R_1, R_2 such that $f(R_1) = f(R_2) = x$, and such that $d^{\text{STARC}}_{\tau,\gamma}(R_1, R_2) > 1 - \epsilon$. This means that the lower diameter of $\text{Am}(f)$ under $d^{\text{STARC}}_{\tau,\gamma}$ is 1. Since 1 is the maximal distance under $d^{\text{STARC}}_{\tau,\gamma}$, we also have that the upper diameter of $\text{Am}(f)$ under $d^{\text{STARC}}_{\tau,\gamma}$ is 1. □

We also need the following lemma:

Lemma 83. *If f_{τ_1} is invariant to S'-redistribution with τ_1, and $\tau_1 \neq \tau_2$, then for all γ and all rewards R, there exists a reward $R^\dagger$ such that $R^\dagger$ is non-trivial under τ_2 and γ, and $f_{\tau_1}(R) = f_{\tau_1}(R + \alpha R^\dagger)$ for all $\alpha \in \mathbb{R}$.*

Proof. Let $R^\dagger$ be a reward such that $\mathbb{E}_{S' \sim \tau_1}\left[R^\dagger(s, a, S')\right] = 0$ for all s, a, and such that $\mathbb{E}_{S' \sim \tau_2}\left[R^\dagger(s, a, S')\right] = 1$ for some s, a such that $\tau_1(s, a) \neq \tau_2(s, a)$. Lemma 80 implies that such a reward function exists. Note that $R^\dagger$ is non-trivial under τ_2 and γ (there is a policy whose value under $R^\dagger$, τ_2, and γ is 0, and a policy whose value is greater than 0). Moreover, for all R and all α, we have that R and $R + \alpha R^\dagger$ differ by S'-redistribution (with τ_1), and so $f_{\tau_1}(R) = f_{\tau_1}(R + \alpha R^\dagger)$. □

Using this lemma, we can now derive a quantitative result. Recall that $d^{\text{STARC}}_{\tau,\gamma}$ is the STARC metric described in Definition 56:

Theorem 84. *If f_{τ_1} is invariant to S'-redistribution with τ_1, and $\tau_1 \neq \tau_2$, then the lower and upper diameter of $\text{Am}(f_{\tau_1})$ under $d^{\text{STARC}}_{\tau_2,\gamma}$ is 1.*

Proof. Immediate from Lemma 82 and 83. □

Note that 1 is the maximal distance that is possible under $d^{\text{STARC}}_{\tau,\gamma}$. This result may be surprising; if $\tau_1 \approx \tau_2$, then one might expect that a reward function that is learnt under τ_1 should be guaranteed to be mostly accurate under τ_2. Note also that Theorem 84 applies for any two transition functions τ_1, τ_2 such that $\tau_1 \neq \tau_2$ *in any state*; it is not required that $\tau_1 \neq \tau_2$ in *every* state. At the end of this section, we will provide an intuitive explanation of Theorem 84.

We will next show that any behavioural model that is invariant to potential shaping is unable to guarantee transfer learning to a different discount factor γ. We

say that τ is *trivial* if for each $s \in \mathcal{S}$, $\tau(s,a) = \tau(s,a')$ for all $a, a' \in \mathcal{A}$. Moreover, we say that a state s is *controllable* relative to a transition function τ, initial state distribution μ_0, and discount γ, if there exist two policies π, π' such that

$$\sum_{t=0}^{\infty} \gamma^t \mathbb{P}_{\xi\sim\pi}(S_t = s) \neq \sum_{t=0}^{\infty} \gamma^t \mathbb{P}_{\xi\sim\pi'}(S_t = s).$$

In other words, a state is controllable if the agent can influence how often it visits that state in expectation. We note the following:

Lemma 85. *For any μ_0, γ, and τ, there exists a controllable state if and only if τ is non-trivial.*

Proof. It is straightforward to see that if τ is trivial then there are no controllable states. For the other direction, suppose there are no controllable states. Let a "state-valued" reward function be a reward function R such that for each $s \in \mathcal{S}$, we have that $R(s,a_1,s_1) = R(s,a_2,s_2)$ for all $s_1, s_2 \in \mathcal{S}, a_1, a_2 \in \mathcal{A}$. Given a state-valued reward R, let $\vec{R} \in \mathbb{R}^{|S|}$ be the vector such that $\vec{R}[s]$ is the reward that R assigns to transitions leaving s. Moreover, given a policy π, let T^π be the $|\mathcal{S}| \times |\mathcal{S}|$-dimensional transition matrix that describes the transitions of π under τ, so that $T^\pi[s,s'] = \mathbb{P}_{A\sim\pi(s),S'\sim\tau(s,A)}(S' = s')$, and let $\vec{V^\pi} \in \mathbb{R}^{|S|}$ be the vector such that $\vec{V^\pi}[s] = V^\pi(s)$. Using the Bellman equation for V^π (Equation 2.1), we now have that that:

$$\begin{aligned}
\vec{V^\pi} &= \vec{R} + \gamma T^\pi \vec{V^\pi} \\
\vec{V^\pi} - \gamma T^\pi \vec{V^\pi} &= \vec{R} \\
(I - \gamma T^\pi)\vec{V^\pi} &= \vec{R} \\
\vec{V^\pi} &= (I - \gamma T^\pi)^{-1}\vec{R}
\end{aligned}$$

To see that $(I - \gamma T^\pi)$ always is invertible, note that the identity $(I - \gamma T^\pi)\vec{V^\pi} = \vec{R}$ implies that we, for any value function V^π, can find a state-valued reward function R such that V^π is the value function for R. Moreover, via the Bellman optimality equation (Equation 2.1), we have that we, for any state-valued reward function R, can find a value function V^π such that V^π is the value function for R. There

is therefore a one-to-one correspondence between value functions and state-valued reward functions, and so $(I - \gamma T^\pi)$ must be invertible.

Next, note that if there are no controllable states, and R is state-valued, then every policy π has the same value function V^π. This means that

$$(I - \gamma T^\pi)^{-1}\vec{R} = (I - \gamma T^{\pi'})^{-1}\vec{R}$$

for all policies π, π'. Next, since R was chosen arbitrarily, this identity must hold for all state-valued reward functions (i.e., all vectors in $\mathbb{R}^{|\mathcal{S}|}$), which means that

$$(I - \gamma T^\pi)^{-1} = (I - \gamma T^{\pi'})^{-1}$$

for all policies π, π'. From this, it follows that $T^\pi = T^{\pi'}$ for all π, π', which in turn implies that the transition function τ must be trivial. □

We can now state the following result:

Theorem 86. *If f_{γ_1} is invariant to potential shaping with γ_1, $\gamma_1 \neq \gamma_2$, and τ is non-trivial, then we have that* $\mathrm{Am}(f_{\gamma_1}) \not\trianglelefteq \mathrm{OPT}_{\tau,\gamma_2}$.

Proof. As per Lemma 85, if τ is non-trivial then there is a state s that is controllable relative to τ and γ_2 (and any μ_0 under which all states are reachable). Let $\Phi_x : \mathcal{S} \to \mathbb{R}$ be the potential function given by $\Phi_x(s) = X$ and $\Phi_x(s') = 0$ for all $s' \neq s$, where $X \in \mathbb{R}$ and $X \neq 0$, and let R be the reward function given by

$$R(s, a, s') = \gamma_1 \cdot \Phi_x(s') - \Phi_x(s).$$

Now R_0 and R differ by potential shaping with γ_1, where R_0 is the reward function that is zero everywhere, which means that $f^{\gamma_1}(R_0) = f^{\gamma_1}(R)$. Let J be the policy value function of R, evaluated under τ, γ_2, and some initial state distribution μ_0 under which all states are reachable. Moreover, given a policy π, let

$$n^\pi = \sum_{t=0}^{\infty} \gamma_2^t \mathbb{P}_{\xi\sim\pi}(S_{t+1} = s),$$
$$x^\pi = \sum_{t=0}^{\infty} \gamma_2^t \mathbb{P}_{\xi\sim\pi}(S_t = s).$$

We then have that $J(\pi) = X \cdot (\gamma_1 n^\pi - x^\pi)$. Let p denote $\mu_0(s)$. If $\gamma_1 = \gamma_2$ then we know that $J(\pi) = -X \cdot p$ (Proposition 29), which gives that

$$\begin{aligned} X \cdot (\gamma_2 n^\pi - x^\pi) &= -X \cdot p \\ \gamma_2 n^\pi - x^\pi &= -p \\ x^\pi &= \gamma_2 n^\pi + p \end{aligned}$$

By plugging this into the above, and rearranging, we obtain

$$J(\pi) = X n^\pi (\gamma_1 - \gamma_2) - pX.$$

Moreover, if s is controllable then there are π_1, π_2 such that $n^{\pi_1} \neq n^{\pi_2}$, which means that $J(\pi_1) \neq J(\pi_2)$. Thus R is not trivial under τ and γ_2. Since R_0 is trivial under τ and γ_2, this means that $R \not\equiv_{\mathrm{OPT}_{\tau,\gamma_2}} R_0$. Thus, there exists reward functions R, R_0 such that $f_{\gamma_1}(R) = f_{\gamma_1}(R_0)$ but $R \not\equiv_{\mathrm{OPT}_{\tau,\gamma_2}} R_0$, which means that $\mathrm{Am}(f^{\gamma_1}) \not\trianglelefteq \mathrm{OPT}_{\tau,\gamma_2}$. □

Note that if τ is trivial, then there can never be any situations where the agent has to decide between obtaining a smaller reward sooner or a greater reward later, which means that the discount factor has no impact on which policies are optimal. This requirement is therefore necessary, although it is very mild. We can also extend this result to a stronger statement, expressed in terms of Definition 6:

Theorem 87. *If f_{γ_1} is invariant to potential shaping with γ_1, $\gamma_1 \neq \gamma_2$, and τ is non-trivial, then the lower and upper diameter of $\mathrm{Am}(f_{\gamma_1})$ under $d^{\mathrm{STARC}}_{\tau,\gamma_2}$ is 1.*

Proof. In the proof of Theorem 86 we show that if $\gamma_1 \neq \gamma_2$ and τ is non-trivial, then there exists a reward $R^\dagger$ such that R and $R + \alpha R^\dagger$ differ by potential shaping with γ_1 (for all R and all $\alpha \in \mathbb{R}$), and such that $R^\dagger$ is non-trivial under γ_2. We can thus apply Lemma 82. □

Again, recall that 1 is the maximal distance that is possible under $d^{\mathrm{STARC}}_{\tau,\gamma}$. This result may also be surprising; if $\gamma_1 \approx \gamma_2$, then one might expect that a reward function that is learnt under γ_1 should be guaranteed to be mostly accurate under γ_2. Before moving on, let us therefore provide an intuitive explanation for these results.

Let us start with Theorem 84. Suppose we have a simple $N \times N$ gridworld environment, as illustrated in Figure 5.2. We assume that the agent has four actions, `up`, `down`, `left`, and `right`. We assume that τ_2 is deterministic, so that if the agent takes action `up`, then it moves one step up, etc. Moreover, we assume that τ_1 is "slippery", so that if the agent takes action `up`, then it moves up, up-left, and up-right with equal probability, and that if it takes action `right`, then it moves right, up-right, and down-right with equal probability, etc. For simplicity, we will also assume that the environment wraps around itself (like a torus), so that if the agent moves up from the top of the environment, then it ends up at the bottom, and so on.

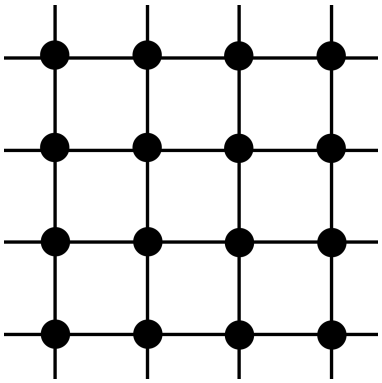
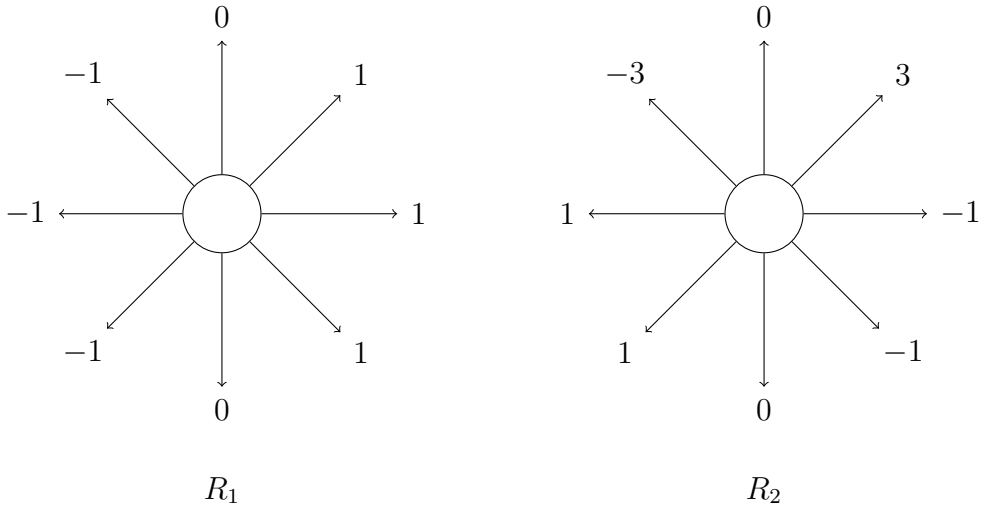


Figure 5.2: A simple illustration of a gridworld environment.

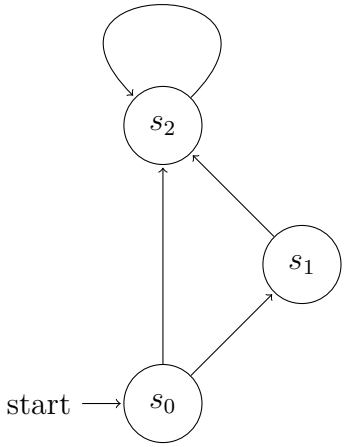
Now suppose that R_1 and R_2 reward each transition (s, a, s') depending on the relative location of s and s', according to the following schemas:



R_1 R_2

These two reward functions are equivalent under τ_1, and give the agent 1 reward for going right, -1 for going left, and 0 for going up or down. However, under τ_2, they are opposites; R_1 rewards the agent for going right, and R_2 rewards the agent for going left. Thus, if we observe a policy computed under τ_1, then we will not be able to distinguish between R_1 and R_2, even though they induce very different behaviour under τ_2. For this reason, it is difficult to obtain guarantees for transfer learning in IRL.

Let us next explain Theorem 87. Consider a simple environment with three states s_0, s_1, s_2, where s_0 is the initial state, and where the agent can choose to either go directly from s_0 to s_2, or choose to first visit state s_1:



Let R_1 be any reward function over this environment, and let R_2 be the reward function that we get if we take R_1 and *increase* the reward of going from s_0 to s_1 by $\gamma_1 \cdot X$, and *decrease* the reward of going from s_1 to s_2 by X. Now, the policy order under discounting with γ_1 is completely unchanged. This transformation corresponds to potential shaping where $\Phi(s_1) = X$ and $\Phi(s_0) = \Phi(s_2) = 0$. Therefore, if $f : \mathcal{R} \to \Pi$ is invariant to potential shaping with γ_1, then $f(R_1) = f(R_2)$. However, if we discount with γ_2, then R_1 and R_2 have a different policy order. In particular, the value of going from s_0 to s_1 is changed by $\gamma_1 \cdot X - \gamma_2 \cdot X = (\gamma_1 - \gamma_2) \cdot X \neq 0$. Thus, if the optimal action under R_1 at s_0 is to go to s_1, then by making X sufficiently large or sufficiently small (depending on whether $\gamma_1 > \gamma_2$, or vice versa), then we can create a reward function R_2 for which the optimal action instead is to go to s_2,

and vice versa. Thus, if we observe a policy computed under γ_1, then we will not be able to distinguish between R_1 and R_2, even though they induce different behaviours when discounting with γ_2. This makes it difficult to ensure robust transfer to a new γ.

Intuitively speaking, we can use potential shaping to move reward around in the MDP (so that the agent receives a larger immediate reward at the cost of a lower reward later, or vice versa). However, to cancel out the effect of the discounting, later rewards must be made larger than immediate rewards. If the discount values do not match, then this "compensation" will also not match, leading to a distortion of the policy ordering. Indeed, we can make it so that this distortion dominates the rest of the reward function.

For even when a living body is moved, there is no way opened to our eyes to see the mind, a thing which cannot be seen by the eyes...

— St. Augustine of Hippo, 400.

6

Misspecification With Equivalence Relations

In this section, we present our results about how robust IRL is to misspecified behavioural models, using the formalisation provided by Definition 7. First, we will derive necessary and sufficient conditions that describe all forms of misspecification that are tolerated by the Boltzmann-rational model, the MCE model, and the optimality model. In so doing, we will also define some broader equivalence classes of behavioural models that are internally robust to misspecification, and which include more behavioural models than the standard three. After this, we will discuss how to generalise some of our results to even wider classes of behavioural models, and show that some of our results should be expected to apply with some universality. After this, we will discuss the case where the environment model is misspecified, as well as the issue of transfer learning. Most of our results in this section are expressed in terms of the two equivalence relations $\mathrm{ORD}_{\tau,\gamma}$ and $\mathrm{OPT}_{\tau,\gamma}$ on $\mathcal{R}$, which were introduced in Appendix 4.

6.1 Necessary and Sufficient Conditions

In this section, we will present necessary and sufficient conditions that describe all forms of misspecification that are tolerated by the Boltzmann-rational model,

the MCE model, and the optimality model. Let Π^+ be the set of all policies such that $\pi(a \mid s) > 0$ for all s, a, and let $F_{\tau,\gamma}$ be the set of all functions $f_{\tau,\gamma} : \mathcal{R} \to \Pi^+$ that, given R, returns a policy π which satisfies

$$\text{argmax}_{a \in \mathcal{A}} \pi(a \mid s) = \text{argmax}_{a \in \mathcal{A}} Q^\star(s, a),$$

where $Q^\star$ is the optimal Q-function for R under τ and γ. In other words, $F_{\tau,\gamma}$ is the set of functions that generate policies which take each action with positive probability, and that take the optimal actions with the highest probability. This class is quite large, and includes e.g. Boltzmann-rational policies (for any β), but it does not include optimal policies (since they do not take all actions with positive probability) or MCE policies (since they may take suboptimal actions with high probability).

Theorem 88. *Let $f_{\tau,\gamma} \in F_{\tau,\gamma}$ be surjective onto Π^+. Then $f_{\tau,\gamma}$ is $\text{OPT}_{\tau,\gamma}$-robust to misspecification with g if and only if $g \in F_{\tau,\gamma}$ and $g \neq f_{\tau,\gamma}$.*

Proof. Let $f_{\tau,\gamma} \in F_{\tau,\gamma}$ be surjective onto Π^+. By definition, we have that $f_{\tau,\gamma}$ is $\text{OPT}_{\tau,\gamma}$-robust to misspecification with g if and only if $\text{Am}(f_{\tau,\gamma}) \preceq \text{OPT}_{\tau,\gamma}$, $g \neq f_{\tau,\gamma}$, $\text{Im}(g) \subseteq \text{Im}(f)$, and if $f_{\tau,\gamma}(R_1) = g(R_2)$ then R_1 and R_2 have the same optimal policies under τ and γ.

Since $f_{\tau,\gamma} \in F_{\tau,\gamma}$, we have that for all R,

$$\text{argmax}_{a \in \mathcal{A}} f_{\tau,\gamma}(R)(a \mid s) = \text{argmax}_{a \in \mathcal{A}} Q^\star(s, a).$$

Moreover, R_1 and R_2 have the same optimal policies under τ and γ if and only if $\text{argmax}_{a \in \mathcal{A}} Q_1^\star(s, a) = \text{argmax}_{a \in \mathcal{A}} Q_2^\star(s, a)$ under τ and γ. Thus, if $f_{\tau,\gamma}(R_1) = f_{\tau,\gamma}(R_2)$ then $R_1 \equiv_{\text{OPT}_{\tau,\gamma}} R_2$, and so $\text{Am}(f_{\tau,\gamma}) \preceq \text{OPT}_{\tau,\gamma}$.

Let $g \in F_{\tau,\gamma}$ and $g \neq f_{\tau,\gamma}$. Since g is a function $\mathcal{R} \to \Pi^+$, and since $f_{\tau,\gamma}$ is surjective onto Π^+, we have that $\text{Im}(g) \subseteq \text{Im}(f_{\tau,\gamma})$. Next, by the same argument as above, if $f_{\tau,\gamma}(R_1) = g(R_2)$ then $\text{argmax}_{a \in \mathcal{A}} Q_1^\star(s, a) = \text{argmax}_{a \in \mathcal{A}} Q_2^\star(s, a)$, which implies that $R_1 \equiv_{\text{OPT}_{\tau,\gamma}} R_2$. Thus $f_{\tau,\gamma}$ is $\text{OPT}_{\tau,\gamma}$-robust to misspecification with g.

Next, suppose $f_{\tau,\gamma}$ is $\text{OPT}_{\tau,\gamma}$-robust to misspecification with g. This means that $\text{Im}(g) \subseteq \text{Im}(f)$, that $f_{\tau,\gamma} \neq g$, and that if $f_{\tau,\gamma}(R_1) = g(R_2)$ then $R_1 \equiv_{\text{OPT}_{\tau,\gamma}} R_2$.

First, note that $\text{Im}(g) \subseteq \text{Im}(f)$ implies that g is a function $\mathcal{R} \to \Pi^+$. Next, let R_1 be an arbitrary reward function, and let R_2 be a reward function such that $f_{\tau,\gamma}(R_2) = g(R_1)$. Since $\text{Im}(g) \subseteq \text{Im}(f)$, we have that such a reward function R_2 must exist. Next, since $f_{\tau,\gamma}(R_2) = g(R_1)$,

$$\text{argmax}_{a \in \mathcal{A}} f_{\tau,\gamma}(R_2)(a \mid s) = \text{argmax}_{a \in \mathcal{A}} g(R_1)(a \mid s).$$

Moreover, we have that $R_1 \equiv_{\text{OPT}_{\tau,\gamma}} R_2$, since $f_{\tau,\gamma}$ is $\text{OPT}_{\tau,\gamma}$-robust to misspecification with g. This means that

$$\text{argmax}_{a \in \mathcal{A}} Q_1^\star(s,a) = \text{argmax}_{a \in \mathcal{A}} Q_2^\star(s,a).$$

Now, since $f_{\tau,\gamma} \in F_{\tau,\gamma}$, we have that

$$\text{argmax}_{a \in \mathcal{A}} f_{\tau,\gamma}(R_2)(a \mid s) = \text{argmax}_{a \in \mathcal{A}} Q_2^\star(s,a).$$

By transitivity, this implies that

$$\text{argmax}_{a \in \mathcal{A}} g(R_1)(a \mid s) = \text{argmax}_{a \in \mathcal{A}} Q_1^\star(s,a).$$

Since R_1 was chosen arbitrarily, this must hold for all R_1. Thus $g \in F_{\tau,\gamma}$. □

Boltzmann-rational policies are surjective onto Π^+. To see this, note that if a policy π takes each action with positive probability, then its action probabilities are always the softmax of some Q-function, and any Q-function corresponds to some reward function (via Equation 2.4). Therefore, Theorem 88 exactly characterises all forms of misspecification to which the Boltzmann-rational model is $\text{OPT}_{\tau,\gamma}$-robust. Specifically, $b_{\tau,\gamma,\beta}$ is $\text{OPT}_{\tau,\gamma}$-robust to misspecification with g if and only if $g(R)$ always is a policy that takes each action with positive probability, and takes the optimal actions with the highest probability. This includes Boltzmann-rational policies with different temperature parameters than β. It also includes the policies that with probability $1-\epsilon$ take an optimal action, and with probability ϵ take a random action (for $\epsilon \in (0, 0.5]$), and so on.

Let us briefly comment on the requirement that $\pi(a \mid s) > 0$, which corresponds to the condition that $\text{Im}(g) \subseteq \text{Im}(f)$ in Definition 7. If a learning algorithm $\mathcal{L}$

is based on a model $f : \mathcal{R} \to \Pi^+$ then it assumes that the observed policy takes each action with positive probability in every state. What happens if such an algorithm $\mathcal{L}$ is given data from a policy that takes some action with probability 0? This depends on $\mathcal{L}$, but for most sensible algorithms the result should simply be that $\mathcal{L}$ assumes that (or acts as if) those actions are taken with a positive but low probability. This means that it should be possible to drop the requirement that $\pi(a \mid s) > 0$ for many reasonable learning algorithms $\mathcal{L}$.

We next consider the misspecification to which the Boltzmann-rational model is $\mathrm{ORD}_{\tau,\gamma}$-robust. Let $\psi : \mathcal{R} \to \mathbb{R}^+$ be any function from reward functions to positive real numbers, and let $b_{\tau,\gamma,\psi} : \mathcal{R} \to \Pi^+$ be the function that, given R, returns the Boltzmann-rational policy with temperature $\psi(R)$ given transition function τ and discount γ. Moreover, let $B_{\tau,\gamma} = \{b_{\tau,\gamma,\psi} : \psi \in \mathcal{R} \to \mathbb{R}^+\}$ be the set of all such functions $b_{\tau,\gamma,\psi}$. This set includes Boltzmann-rational policies; just let ψ return a constant β for all R.

Theorem 89. *For any $\beta > 0$, $b_{\tau,\gamma,\beta}$ is $\mathrm{ORD}_{\tau,\gamma}$-robust to misspecification with g if and only if $g \in B_{\tau,\gamma}$ and $g \neq b_{\tau,\gamma,\beta}$.*

Proof. As per Theorem 73, $\mathrm{Am}(b_{\tau,\gamma,\beta})$ is characterised by $\mathrm{PS}_\gamma \odot S'\mathrm{R}_\tau$, and as per Theorem 40, $\mathrm{ORD}_{\tau,\gamma}$ is characterised by $\mathrm{PS}_\gamma \odot S'\mathrm{R}_\tau \odot \mathrm{LS}$. Hence $\mathrm{Am}(b_{\tau,\gamma,\beta}) \preceq \mathrm{ORD}_{\tau,\gamma}$, which means that Lemma 15 implies that $b_{\tau,\gamma,\beta}$ is $\mathrm{ORD}_{\tau,\gamma}$-robust to misspecification with g if and only if $g \neq b_{\tau,\gamma,\beta}$, and there exists a $t \in \mathrm{PS}_\gamma \odot S'\mathrm{R}_\tau \odot \mathrm{LS}$ such that $g = b_{\tau,\gamma,\beta} \circ t$.

For the first direction, assume that there exists a $t \in \mathrm{PS}_\gamma \odot S'\mathrm{R}_\tau \odot \mathrm{LS}$ such that $g = b_{\tau,\gamma,\beta} \circ t$ and $g \neq b_{\tau,\gamma,\beta}$. Now $b_{\tau,\gamma,\beta}(R)$ is the policy given by

$$b_{\tau,\gamma,\beta}(R)(a \mid s) = \frac{\exp \beta A_R(s,a)}{\sum_{a \in \mathcal{A}} \exp \beta A_R(s,a)}.$$

where A_R is the optimal advantage function of R under τ and γ. If $g(R) = b_{\tau,\gamma,\beta} \circ t(R)$ for some $t \in \mathrm{PS}_\gamma \odot \mathrm{LS} \odot S'\mathrm{R}_\tau$, then we have that

$$\begin{aligned} g(R)(a \mid s) &= \frac{\exp \beta A_{t(R)}(s,a)}{\sum_{a \in \mathcal{A}} \exp \beta A_{t(R)}(s,a)} \\ &= \frac{\exp \beta c_R A_R(s,a)}{\sum_{a \in \mathcal{A}} \exp \beta c_R A_R(s,a)}, \end{aligned}$$

where c_R is the linear scaling factor that t applies to R. Note that the advantage function A is preserved by both potential shaping and S'-redistribution (Lemma 70). Now let $\psi(R) = \beta \cdot c_R$, and we can see that $g = b_{\tau,\gamma,\psi} \in B_{\tau,\gamma}$. Thus, if $b_{\tau,\gamma,\beta}$ is $\mathrm{ORD}_{\tau,\gamma}$-robust to misspecification with g, then $g \in B_{\tau,\gamma}$ and $g \neq b_{\tau,\gamma,\psi}$.

For the other direction, assume that $g \in B_{\tau,\gamma}$ and $g \neq b_{\tau,\gamma,\beta}$. Since $g \in B_{\tau,\gamma}$, there is a function $\psi : \mathcal{R} \to \mathbb{R}^+$ such that $g(R)$ is the policy given by applying a softmax function with temperature $\psi(R)$ to the optimal advantage function of R. Now let $t \in \mathrm{LS}$ be the function that scales each $R \in \mathcal{R}$ by a factor of $\psi(R)/\beta$, and we can see that $g = b_{\tau,\gamma,\beta} \circ t$. This completes the proof. □

Theorem 89 thus says that the Boltzmann-rational model is $\mathrm{ORD}_{\tau,\gamma}$-robust to misspecification of the temperature parameter β, but not to any other form of misspecification (with the only complication being that the misspecification of β is allowed to depend arbitrarily on the underlying reward function). We next turn our attention to optimal policies.

Theorem 90. *For each $o \in \mathcal{O}_{\tau,\gamma}$, we have that $\mathrm{Am}(o) \not\preceq \mathrm{ORD}_{\tau,\gamma}$, unless $|\mathcal{S}| = 1$ and $|\mathcal{A}| = 2$. The only function $o \in \mathcal{O}_{\tau,\gamma}$ such that $\mathrm{Am}(o) \preceq \mathrm{OPT}_{\tau,\gamma}$ is $o^\star_{\tau,\gamma}$, but there is no function g such that $o^\star_{\tau,\gamma}$ is $\mathrm{OPT}_{\tau,\gamma}$-robust to misspecification with g.*

Proof. The first part follows from Proposition 77, which says that $\mathrm{Am}(o) \not\preceq \mathrm{ORD}_{\tau,\gamma}$, unless $|\mathcal{S}| = 1$ and $|\mathcal{A}| = 2$. Moreover, $\mathrm{Am}(o^\star_{\tau,\gamma}) = \mathrm{OPT}_{\tau,\gamma}$ (Theorem 75). Therefore, by Lemma 14, there is no function g such that $o^\star_{\tau,\gamma}$ is $\mathrm{OPT}_{\tau,\gamma}$-robust to misspecification with g. Finally, Theorem 79 says that if $o \in \mathcal{O}_{\tau,\gamma}$ but $o \neq o^\star_{\tau,\gamma}$, then $\mathrm{Am}(o) \not\preceq \mathrm{OPT}_{\tau,\gamma}$. □

This essentially means that the optimality model is not robust to any form of misspecification (regardless of whether that is measured using $\mathrm{ORD}_{\tau,\gamma}$ or $\mathrm{OPT}_{\tau,\gamma}$). We finally turn our attention to causal entropy maximising policies. As before, let $\psi : \mathcal{R} \to \mathbb{R}^+$ be any function from reward functions to positive real numbers, and let $c_{\tau,\gamma,\psi} : \mathcal{R} \to \Pi^+$ be the function that, given R, returns the MCE policy with weight $\psi(R)$ given τ and γ. Furthermore, let $C_{\tau,\gamma} = \{c_{\tau,\gamma,\psi} : \psi \in \mathcal{R} \to \mathbb{R}^+\}$ be the set of all such functions $c_{\tau,\gamma,\psi}$. Moreover, as usual, we let $c_{\tau,\gamma,\alpha} : \mathcal{R} \to \Pi^+$ be the

function that, given R, returns the MCE policy with weight α given τ and γ. Also note that $c_{\tau,\gamma,\alpha} \in C_{\tau,\gamma}$ for each α, since we may let ψ return a constant α for all R.

Theorem 91. *For any $\alpha > 0$, we have that $c_{\tau,\gamma,\alpha}$ is $\mathrm{ORD}_{\tau,\gamma}$-robust to misspecification with g if and only if $g \in C_{\tau,\gamma}$ and $g \neq c_{\tau,\gamma,\psi}$.*

Proof. As per Theorem 74, $\mathrm{Am}(c_{\tau,\gamma,\alpha})$ is characterised by $\mathrm{PS}_\gamma \odot S'\mathrm{R}_\tau$, and as per Theorem 40, $\mathrm{ORD}_{\tau,\gamma}$ is characterised by $\mathrm{PS}_\gamma \odot S'\mathrm{R}_\tau \odot \mathrm{LS}$. Hence $\mathrm{Am}(c_{\tau,\gamma,\alpha}) \preceq \mathrm{ORD}_{\tau,\gamma}$, which means that Lemma 15 implies that $c_{\tau,\gamma,\alpha}$ is $\mathrm{ORD}_{\tau,\gamma}$-robust to misspecification with g if and only if $g \neq c_{\tau,\gamma,\alpha}$, and there exists a $t \in \mathrm{PS}_\gamma \odot S'\mathrm{R}_\tau \odot \mathrm{LS}$ such that $g = c_{\tau,\gamma,\alpha} \circ t$.

For the first direction, assume that $g \neq c_{\tau,\gamma,\alpha}$, and that there exists a $t \in \mathrm{PS}_\gamma \odot S'\mathrm{R}_\tau \odot \mathrm{LS}$ such that $g = c_{\tau,\gamma,\alpha} \circ t$. Recall that $c_{\tau,\gamma,\alpha}(R)$ is the unique policy that maximises the maximal causal entropy objective;

$$J_R^{\mathrm{MCE}}(\pi) = J_R(\pi) - \alpha \sum_{t=0}^{\infty} \mathbb{E}_{S_t \sim \pi,\tau,\mu_0}[\gamma^t H(\pi(S_t))],$$

where J_R is the policy evaluation function for the reward function R. Therefore, if $g(R) = c_{\tau,\gamma,\psi} \circ t(R)$ then $g(R)$ is the policy

$$\begin{aligned}
&\max_\pi J_{t(R)}^{\mathrm{MCE}}(\pi) \\
=&\max_\pi J_{t(R)}(\pi) - \alpha \sum_{t=0}^{\infty} \mathbb{E}_{S_t \sim \pi,\tau,\mu_0}[\gamma^t H(\pi(S_t))] \\
=&\max_\pi c_R \cdot J_R(\pi) - \alpha \sum_{t=0}^{\infty} \mathbb{E}_{S_t \sim \pi,\tau,\mu_0}[\gamma^t H(\pi(S_t))]
\end{aligned}$$

where c_R is the linear scaling factor that t applies to R. Note that J_R is preserved by S'-redistribution, and potential shaping can only change J_R by inducing a uniform constant shift of J_R for all policies (Proposition 29). Thus linear scaling is the only transformation in $\mathrm{PS}_\gamma \odot S'\mathrm{R}_\tau \odot \mathrm{LS}$ that could affect the MCE objective. Finally, let ψ be the function $\psi(R) = \alpha / c_R$, and we can see that $g = c_{\tau,\gamma,\psi} \in C_{\tau,\gamma}$.

For the other direction, assume that $g \in C_{\tau,\gamma}$ and $g \neq c_{\tau,\gamma,\psi}$. Then there is a function $\psi : \mathcal{R} \to \mathbb{R}^+$ such that $g(R)$ is the unique policy that maximises the MCE objective given by

$$J_R(\pi) - \psi(R) \sum_{t=0}^{\infty} \mathbb{E}_{S_t \sim \pi,\tau,\mu_0}[\gamma^t H(\pi(S_t))].$$

Now let $t \in \text{LS}$ be the function that applies a positive linear scaling factor of $\psi(R)/\alpha$ to each reward function R, and we can see that $g = c_{\tau,\gamma,\psi} \circ t$. Since $t \in \text{LS}$, this completes the other direction, and the proof. □

In other words, the maximal causal entropy model is $\text{ORD}_{\tau,\gamma}$-robust to misspecification of the weight α, but not to any other kind of misspecification (with the only complication being that the misspecification of α is allowed to depend arbitrarily on the underlying reward function). This is similar to what is the case for Boltzmann-rational policies.

Finally, let us briefly discuss the misspecification to which the maximal causal entropy model is $\text{OPT}_{\tau,\gamma}$-robust. Lemma 15 tells us that $c_{\tau,\gamma,\alpha}$ is $\text{OPT}_{\tau,\gamma}$-robust to misspecification with g if $g = c_{\tau,\gamma,\alpha} \circ t$ for some $t \in \text{OPT}_{\tau,\gamma}$. In other words, if $g(R_1) = \pi$ then there must exist an R_2 such that π maximises causal entropy with respect to R_2, and such that R_1 and R_2 have the same optimal policies. It seems hard to express this as an intuitive property of g, so we have refrained from stating this result as a theorem.

6.2 Wider Classes of Policies

At this point, it is worth remarking on the fact that there are several noteworthy parallels between the invariances and the misspecification robustness of $b_{\tau,\gamma,\beta}$ and $c_{\tau,\gamma,\alpha}$. In particular, both are invariant to potential shaping and S'-redistribution, and no other transformations. Moreover, both are defined in terms of a parameter (β or α), and both are $\text{ORD}_{\tau,\gamma}$-robust to misspecification of this parameter, and no other forms of misspecification. Additionally, misspecification of this parameter results in positive linear scaling of the learnt reward function. Is this a coincidence, or should we expect the same result to generalise to a wider class of behavioural models? Before moving on, we will discuss this question in some more depth.

First of all, Q-functions and advantage functions are invariant to S'-redistribution (Lemma 67-71). It is easy to show that value functions and policy evaluation functions, etc, also are invariant to S'-redistribution. This means that any behavioural

model which can be computed via one of these objects also will share this invariance, as per Lemma 4. More generally, since S'-redistribution does not change the expected value of any policy in any state, it is quite natural for a behavioural policy to be invariant to such transformations. It is also quite natural for a behavioural model to be invariant to potential shaping, if that behavioural model uses exponential discounting, considering the properties of potential shaping discussed in Section 4.1.

Lemma 14 tells us that for f to be $\mathrm{ORD}_{\tau,\gamma}$-robust to some forms of misspecification, it has to be the case that f is sensitive to some reward transformations which do not affect the policy order of the reward function. For example, $b_{\tau,\gamma,\beta}$ and $c_{\tau,\gamma,\alpha}$ are sensitive to positive linear scaling, even though this does not affect the policy order. Lemma 15 then tells us that f can be composed with these transformations, to produce the forms of misspecification that f will tolerate. Composing $b_{\tau,\gamma,\beta}$ or $c_{\tau,\gamma,\alpha}$ with positive linear scaling is equivalent to scaling β or α; hence $b_{\tau,\gamma,\beta}$ and $c_{\tau,\gamma,\alpha}$ are $\mathrm{ORD}_{\tau,\gamma}$-robust to such misspecification. But is it reasonable to expect a behavioural model to be sensitive to some order-preserving transformations? We next show that if $f : \mathcal{R} \to \Pi^+$ is continuous, surjective onto Π^+, and satisfies $\mathrm{Am}(f) \preceq \mathrm{ORD}_{\tau,\gamma}$, then f must be sensitive to some such transformations:

Proposition 92. *If $f : \mathcal{R} \to \Pi^+$ is continuous, and $f(R_1) = f(R_2)$ if and only if $R_1 \equiv_{\mathrm{ORD}_{\tau,\gamma}} R_2$, then f is not surjective onto Π^+.*

Proof. Assume for contradiction that $f : \mathcal{R} \to \Pi^+$ is continuous and surjective, and that $f(R_1) = f(R_2)$ if and only if $R_1 \equiv_{\mathrm{ORD}_{\tau,\gamma}} R_2$. Then f is a continuous bijection from $\mathrm{Im}(s^{\mathrm{STARC}}_{\tau,\gamma})$ to Π^+, where $s^{\mathrm{STARC}}_{\tau,\gamma}$ is the standardisation function of $d^{\mathrm{STARC}}_{\tau,\gamma}$ (Definition 56). Moreover, $\mathrm{Im}(s^{\mathrm{STARC}}_{\tau,\gamma})$ is compact (because it is a closed and bounded subset of a finite-dimensional Euclidean space), and Π^+ is Hausdorff. It thus follows that f is a homeomorphism. This is a contradiction, since Π^+ is not homeomorphic to $\mathrm{Im}(s^{\mathrm{STARC}}_{\tau,\gamma})$. For example, $\mathrm{Im}(s^{\mathrm{STARC}}_{\tau,\gamma})$ contains an isolated point, which Π^+ does not. □

Thus, if we want f to be both continuous and surjective onto Π^+, then there must either be some order-preserving reward transformations to which f is not

invariant (i.e. $\mathrm{Am}(f) \prec \mathrm{ORD}_{\tau,\gamma}$), or f must be invariant to some transformations which are not order preserving (i.e. $\mathrm{Am}(f) \not\preceq \mathrm{ORD}_{\tau,\gamma}$). The latter case would imply that f violates condition 3 in Definition 7, and thus that f is not robust to any misspecification. In the former case, which transformations would it be reasonable to pick? We next show that linear scaling is a natural choice:

Proposition 93. *If $f : \mathcal{R} \to \Pi^+$ is continuous and invariant to positive linear scaling, then $f(R_1) = f(R_2)$ for all R_1, R_2.*

Proof. This is straightforward. Suppose $f : \mathcal{R} \to \Pi^+$ is continuous and invariant to positive linear scaling. Let R_1, R_2 be two arbitrary reward functions, and consider a sequence c_t where $c_t > 0$, but $c_t \to 0$ as $t \to \infty$. Next, consider the sequence given by $c_t \cdot R_1$. Since $c_t \cdot R_1 \to R_0$ as $t \to \infty$, and since f is continuous, we have that $f(c_t \cdot R_1) \to f(R_0)$ as $t \to \infty$. Moreover, since f is invariant to positive linear scaling, we have that $f(c_t \cdot R_1) = f(R_1)$ for all c_t. This implies that $f(R_1) = f(R_0)$. By an analogous argument, we also have that $f(R_2) = f(R_0)$, and hence that $f(R_1) = f(R_2)$. □

Of course, if $f(R_1) = f(R_2)$ for all R_1, R_2, then f is completely trivial. Thus, a continuous behavioural model f should not be (everywhere) invariant to positive linear scaling. Of course, it could be the case that f is sensitive to positive linear scaling in the vicinity of R_0, but otherwise invariant to positive linear scaling, although this seems somewhat unnatural. This then suggests that if a behavioural model $f : \mathcal{R} \to \Pi^+$ is continuous, surjective, and satisfies $\mathrm{Am}(f) \preceq \mathrm{ORD}_{\tau,\gamma}$, then it is natural for f to be sensitive to positive linear scaling, in which case f can be composed with positive linear scaling to produce forms of misspecification to which f is robust.[1] This is exemplified by $b_{\tau,\gamma,\beta}$ and $c_{\tau,\gamma,\alpha}$, and the above argument suggests that we should expect a similar result to hold for many other behavioural models.

[1]Very roughly and informally, a set of reward functions in which no two reward functions share the same policy order will be one dimension short of being able to cover the set of all policies. Therefore, if f *does* cover all policies, then it must be sensitive to one "dimension" of order-preserving transformations. This, in turn, translates to one "dimension" of misspecification to which f is $\mathrm{ORD}_{\tau,\gamma}$-robust.

6.3 Misspecified Parameters

A behavioural model will typically be parameterised by a γ or τ, implicitly or explicitly. In this section, we explore what happens if these parameters are misspecified. We show that a wide class of behavioural models lack robustness to this type of misspecification.

Theorems 88-91 already tell us that the standard behavioural models are not ($\mathrm{ORD}_{\tau,\gamma}$ or $\mathrm{OPT}_{\tau,\gamma}$) robust to misspecified γ or τ, since the sets $F_{\tau,\gamma}$, $B_{\tau,\gamma}$, and $C_{\tau,\gamma}$, all are parameterised by γ and τ. We will generalise this further. First, we note that any behavioural model that is invariant to S'-redistribution will lack robustness to a misspecified τ. Recall Theorem 81; if f_{τ_1} is invariant to S'-redistribution with τ_1, and $\tau_1 \neq \tau_2$, then we have that $\mathrm{Am}(f_{\tau_1}) \not\trianglelefteq \mathrm{OPT}_{\tau_2,\gamma}$. Using this, we can prove the following:

Theorem 94. *If f_τ is invariant to S'-redistribution with τ, and $\tau_1 \neq \tau_2$, then f_{τ_1} is not $\mathrm{OPT}_{\tau_3,\gamma}$-robust to misspecification with f_{τ_2} for any τ_3 or γ.*

Proof. Suppose for contradiction that f_{τ_1} is $\mathrm{OPT}_{\tau_3,\gamma}$-robust to misspecification with f_{τ_2}. If $\tau_1 \neq \tau_2$, then $\tau_3 \neq \tau_1$, or $\tau_3 \neq \tau_2$, or both. Theorem 81 then implies that $\mathrm{Am}(f_{\tau_1}) \not\trianglelefteq \mathrm{OPT}_{\tau_3,\gamma}$, or $\mathrm{Am}(f_{\tau_2}) \not\trianglelefteq \mathrm{OPT}_{\tau_3,\gamma}$, or both. The former violates condition 3 in Definition 7, and the latter violates Lemma 12. Thus f_{τ_1} cannot be $\mathrm{OPT}_{\tau_3,\gamma}$-robust to misspecification with f_{τ_2}. □

Recall that all of the three standard behavioural models are invariant to S'-redistribution, and thus subject to Theorem 94. More generally, since S'-redistribution does not change the expected value of any policy in any state, it is quite natural for a behavioural model to be invariant to S'-redistribution. We should therefore expect Theorem 94 to apply very broadly. Also recall that if f_{τ_1} is not $\mathrm{OPT}_{\tau_3,\gamma}$-robust to misspecification with f_{τ_2}, then it is also not $\mathrm{ORD}_{\tau_3,\gamma}$-robust to misspecification with f_{τ_2}.

Similarly, we can show that a behavioural model which is invariant to potential shaping will not be robust to misspecification of γ. Recall Theorem 86; if f_{γ_1}

is invariant to potential shaping with γ_1, $\gamma_1 \neq \gamma_2$, and τ is non-trivial, then $\mathrm{Am}(f_{\gamma_1}) \not\trianglelefteq \mathrm{OPT}_{\tau,\gamma_2}$. This implies the following:

Theorem 95. *If f_γ is invariant to potential shaping with γ, $\gamma_1 \neq \gamma_2$, and τ is non-trivial, then f_{γ_1} is not $\mathrm{OPT}_{\tau,\gamma_3}$-robust to misspecification with f_{γ_2} for any γ_3.*

Proof. Suppose for contradiction that f_{γ_1} is $\mathrm{OPT}_{\tau,\gamma_3}$-robust to misspecification with f_{γ_2}. If $\gamma_1 \neq \gamma_2$, then $\gamma_3 \neq \gamma_1$, or $\gamma_3 \neq \gamma_2$, or both. Theorem 86 then implies that $\mathrm{Am}(f_{\gamma_1}) \not\trianglelefteq \mathrm{OPT}_{\tau,\gamma_3}$, or $\mathrm{Am}(f_{\gamma_2}) \not\trianglelefteq \mathrm{OPT}_{\tau,\gamma_3}$, or both. The former violates condition 3 in Definition 7, and the latter violates Lemma 12. Thus f_{γ_1} cannot be $\mathrm{OPT}_{\tau,\gamma_3}$-robust to misspecification with f_{γ_2}. □

In other words, if a behavioural model is invariant to S'-redistribution, then that model is not $\mathrm{OPT}_{\tau,\gamma}$-robust (and therefore also not $\mathrm{ORD}_{\tau,\gamma}$-robust) to misspecification of the transition function τ. Similarly, if the behavioural model is invariant to potential shaping, then that model is not $\mathrm{OPT}_{\tau,\gamma}$-robust (and therefore also not $\mathrm{ORD}_{\tau,\gamma}$-robust) to misspecification of the discount parameter γ. These results should apply to most behavioural models. For example, we can derive the following corollary:

Corollary 96. *Let $f_{\tau,\gamma} : \mathcal{R} \to (\mathcal{S}\times\mathcal{A} \to \mathbb{R})$ be the function that, given a reward R, returns the optimal Q-function $Q^\star$ for R under τ and γ. Suppose $g_{\tau,\gamma} = h \circ f_{\tau,\gamma}$ for some h, and let $\tau_1 \neq \tau_2$. Then $g_{\tau_1,\gamma}$ is not $\mathrm{OPT}_{\tau_3,\gamma}$-robust to misspecification with $g_{\tau_2,\gamma}$ for any τ_3 or γ.*

Proof. By Lemma 68, we have that $f_{\tau,\gamma}$ determines R up to $S'\mathrm{R}_\tau$. Thus, by Theorem 94, we have that $f_{\tau_1,\gamma}$ is not $\mathrm{OPT}_{\tau_3,\gamma}$-robust to misspecification with $f_{\tau_2,\gamma}$. Moreover, $\mathrm{Im}(f_{\tau_1,\gamma}) = \mathrm{Im}(f_{\tau_2,\gamma})$, since for any function $q : \mathcal{S}\times\mathcal{A} \to \mathbb{R}$, we can always find a reward function R such that the optimal Q-function for R is q (via the Bellman optimality equation). We thus apply Lemma 11, and conclude the proof. □

In other words, any policy which can be derived from $Q^\star$ is not robust to misspecification of τ. This is because $Q^\star$ already is invariant to S'-redistribution, and so this follows from Lemma 11 and Theorem 94. Lemma 11 could also be used to derive analogous results for other intermediate reward objects, such as those discussed in Section 5.1.

6.4 Transfer Learning

The equivalence relations we have worked with ($\mathrm{OPT}_{\tau,\gamma}$ and $\mathrm{ORD}_{\tau,\gamma}$) only guarantee that the learnt reward function R_H has the same optimal policies, or ordering of policies, as the true reward $R^\star$ for a given choice of τ and γ. A natural question is what happens if we strengthen this requirement, and demand that R_H has the same optimal policies, or ordering of policies, as $R^\star$, for any choice of τ or γ. We briefly discuss this setting here.

In short, it is impossible to guarantee transfer to any τ or γ. This is already implied by the results in Section 5.4. In particular, if $f_{\tau,\gamma}$ is invariant to S'-redistribution (with τ) and potential shaping (with γ), then

$$\mathrm{Am}(f_{\tau_1,\gamma_1}) \not\trianglelefteq \mathrm{OPT}_{\tau_2,\gamma_2}$$

if either $\tau_1 \neq \tau_2$, or $\gamma_1 \neq \gamma_2$ and τ_2 is non-trivial. Then f_{τ_1,γ_1} will violate condition 3 in Definition 7. Since each of the standard behavioural models are invariant to S'-redistribution and potential shaping, this applies to all of them.

On two occasions I have been asked, 'Pray, Mr. Babbage, if you put into the machine wrong figures, will the right answers come out?' I am not able rightly to apprehend the kind of confusion of ideas that could provoke such a question.

— Charles Babbage, 1864.

7 Misspecification With Metrics

In this section, we present our results about how robust IRL is to misspecified behavioural models, using the formalisation provided by Definition 8. First, we will derive necessary and sufficient conditions that describe all forms of misspecification that are tolerated by the Boltzmann-rational model and the MCE model, and discuss the issue of how to derive similar results for the optimality model. After this, we analyse a particular form of misspecification, which we refer to as *perturbation*, provide necessary and suffucient conditions for a behavioural model to be robust to such misspecification, and show that none of the three main behavioural models meet these conditions. After this, we will discuss the case where the environment model is misspecified, as well as the issue of transfer learning. Section 7.1 is quite dense, but 7.2 and 7.3 both provide more intuitive takeaways.

Our results in this section are expressed in terms of pseudometrics on $\mathcal{R}$. Most of these results apply for any choice of pseudometric, but when we need to select a specific pseudometric, we will use the STARC metric $d^{\text{STARC}}_{\tau,\gamma}$, as specified in Definition 56.

7.1 Necessary and Sufficient Conditions

Recall that if $f : \mathcal{R} \to X$ is a behavioural model such that if $f(R_1) = f(R_2)$ then $d^{\mathcal{R}}(R_1, R_2) = 0$, then we can use Lemma 20 to derive necessary and sufficient conditions for the types of misspecification that f is robust to (as measured by $d^{\mathcal{R}}$). Also recall that if $d^{\mathcal{R}}$ is both sound and complete, then $d^{\mathcal{R}}(R_1, R_2) = 0$ if and only if R_1 and R_2 induce the same ordering of policies (Proposition 45). Moreover, if f is either $b_{\tau,\gamma,\beta}$ or $c_{\tau,\gamma,\alpha}$, then $f(R_1) = f(R_2)$ if and only if R_1 and R_2 differ by potential shaping with γ and S'-redistribution with τ (Theorem 73 and 74), and both of these transformations preserve the policy order under τ and γ (Theorem 40). This means that if $d^{\mathcal{R}}$ is sound and complete, then $b_{\tau,\gamma,\beta}$ and $c_{\tau,\gamma,\alpha}$ satisfy the assumptions for Lemma 20, and so we can use it to characterise the forms of misspecification that these models will tolerate. To do this, we need to find the set T_ϵ of all transformations $t : \mathcal{R} \to \mathcal{R}$ such that $d^{\mathcal{R}}(R, t(R)) \leq \epsilon$ for all R. We thus begin by deriving this set T_ϵ, for the STARC metric $d^{\text{STARC}}_{\tau,\gamma}$:

Theorem 97. *For any $\epsilon < 0.5$, $t : \mathcal{R} \to \mathcal{R}$ satisfies that*

$$d^{\text{STARC}}_{\tau,\gamma}(R, t(R)) \leq \epsilon$$

for all $R \in \mathcal{R}$ if and only if t can be expressed as $t_1 \circ t_2 \circ t_3$ where

$$L_2(R, t_2(R)) \leq L_2(c^{\text{STARC}}_{\tau,\gamma}(R)) \cdot \sin(2 \arcsin(\epsilon))$$

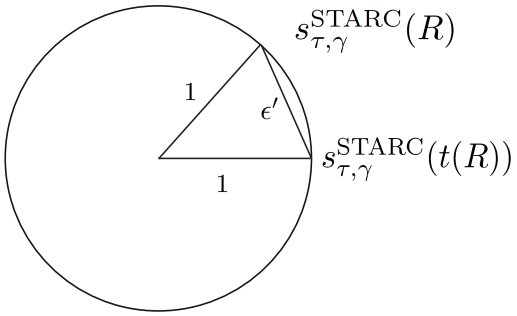
for all R, and where $t_1, t_3 \in S'\text{R}_\tau \bigodot \text{PS}_\gamma \bigodot \text{LS}$.

Proof. For the first direction, suppose $d^{\text{STARC}}_{\tau,\gamma}(R, t(R)) \leq \epsilon$ for all $R \in \mathcal{R}$, and let R be an arbitrarily selected reward function. We will show that it is possible to navigate from R to $t(R)$ using the described transformations.

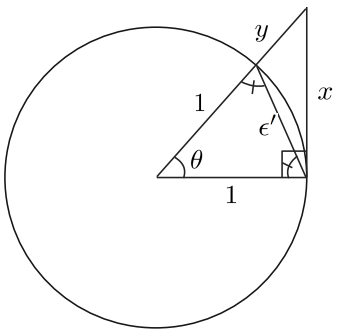
Recall that $d^{\text{STARC}}_{\tau,\gamma}(R, t(R))$ is computed by first applying $c^{\text{STARC}}_{\tau,\gamma}$ to both R and $t(R)$, normalising the resulting vectors, measuring their L_2-distance, and dividing the result by 2. This means that if $d^{\text{STARC}}_{\tau,\gamma}(R, t(R)) < 0.5$, then the L_2-distance between $s^{\text{STARC}}_{\tau,\gamma}(R)$ and $s^{\text{STARC}}_{\tau,\gamma}(t(R))$ is less than 1. Note also that if R is trivial and $t(R)$ is non-trivial, or vice versa, then the L_2-distance between $s^{\text{STARC}}_{\tau,\gamma}(R)$ and

$s^{\text{STARC}}_{\tau,\gamma}(t(R))$ is exactly 1. Thus, either R and $t(R)$ are both trivial, or they are both non-trivial.

If R and $t(R)$ are both trivial, then they differ by some transformation in $\text{PS}_\gamma \bigodot S'\text{R}_\tau$ (as implied by Theorem 40), and so the theorem holds. Next, if R and $t(R)$ are both non-trivial, then $s^{\text{STARC}}_{\tau,\gamma}(R)$ and $s^{\text{STARC}}_{\tau,\gamma}(t(R))$ can be placed in the following diagram, where $\epsilon' \leq 2\epsilon$:



Now, elementary trigonometry tells us that the angle θ between $s^{\text{STARC}}_{\tau,\gamma}(R)$ and $s^{\text{STARC}}_{\tau,\gamma}(t(R))$ is $\theta = 2\arcsin(\epsilon'/2)$. [1] Moreover, suppose we make a right triangle by extending $s^{\text{STARC}}_{\tau,\gamma}(R)$ as follows:



Note that this can be done, since $\epsilon' \leq 2\epsilon < 2 \cdot 0.5 = 1 < \sqrt{2}$. Here elementary trigonometry again tells us that

$$x/(1+y) = \sin(\theta) = \sin(2\arcsin(\epsilon'/2)),$$

[1] This can be seen by bisecting the triangle along the vertex between $s^{\text{STARC}}_{\tau,\gamma}(R)$ and $s^{\text{STARC}}_{\tau,\gamma}(t(R))$, to form two right triangles. Since the hypotenuse is 1, we have that $\sin(\theta/2) = \epsilon'/2$.

or that $x = (1 + y)\sin(2\arcsin(\epsilon'/2))$. This means that we can go from R to $t(R)$ as follows:

1. Apply $c^{\text{STARC}}_{\tau,\gamma}$. Since R and $c^{\text{STARC}}_{\tau,\gamma}(R)$ differ by potential shaping and S'-redistribution, this transformation can be expressed as a combination of potential shaping and S'-redistribution. Call the resulting vector R'.

2. Normalise R', so that its magnitude is 1. This transformation is an instance of positive linear scaling. Call the resulting vector R''.

3. Scale R'' until it forms a right triangle with $s^{\text{STARC}}_{\tau,\gamma}(t(R))$. This transformation is an instance of positive linear scaling. Call the resulting vector R'''.

4. Move from R''' to $s^{\text{STARC}}_{\tau,\gamma}(t(R))$. This will move R''' by a distance equal to $(1 + y)\sin(2\arcsin(\epsilon'/2))$, where $(1 + y) = L_2(R''')$. Moreover, since R''' is in the image of $c^{\text{STARC}}_{\tau,\gamma}$, we have that $R''' = c^{\text{STARC}}_{\tau,\gamma}(R''')$, and so $L_2(R''') = L_2(c^{\text{STARC}}_{\tau,\gamma}(R'''))$. This means that R''' is moved by $L_2(c^{\text{STARC}}_{\tau,\gamma}(R'''))\cdot \sin(2\arcsin(\epsilon'/2))$. Since $0 \leq \epsilon' \leq 2\epsilon < 1 < \sqrt{2}$, and since $\sin(2\arcsin(x/2))$ is growing monotonically on $[0, \sqrt{2}]$ this means that

$$L_2(R''', s^{\text{STARC}}_{\tau,\gamma}(t(R))) \leq L_2(c^{\text{STARC}}_{\tau,\gamma}(R''')) \cdot \sin(2\arcsin(\epsilon)).$$

5. Move from $s^{\text{STARC}}_{\tau,\gamma}(t(R))$ to $c^{\text{STARC}}_{\tau,\gamma}(t(R))$. Since $s^{\text{STARC}}_{\tau,\gamma}(t(R))$ is simply a normalised version of $c^{\text{STARC}}_{\tau,\gamma}(t(R))$, this is an instance of positive linear scaling.

6. Move from $c^{\text{STARC}}_{\tau,\gamma}(t(R))$ to $t(R)$. Since $t(R)$ and $c^{\text{STARC}}_{\tau,\gamma}(t(R))$ differ by potential shaping and S'-redistribution, this transformation can be expressed as a combination of potential shaping and S'-redistribution.

Thus, for an arbitrary reward function R, we can find a series of transformations that fit the given description. This completes the first direction.

For the other direction, suppose t can be expressed as $t_1 \circ t_2 \circ t_3$ where

$$L_2(R, t_2(R)) \leq L_2(c^{\text{STARC}}_{\tau,\gamma}(R)) \cdot \sin(2\arcsin(\epsilon))$$

for all R, and where $t_1, t_3 \in S'\mathrm{R}_\tau \bigodot \mathrm{PS}_\gamma \bigodot \mathrm{LS}$.

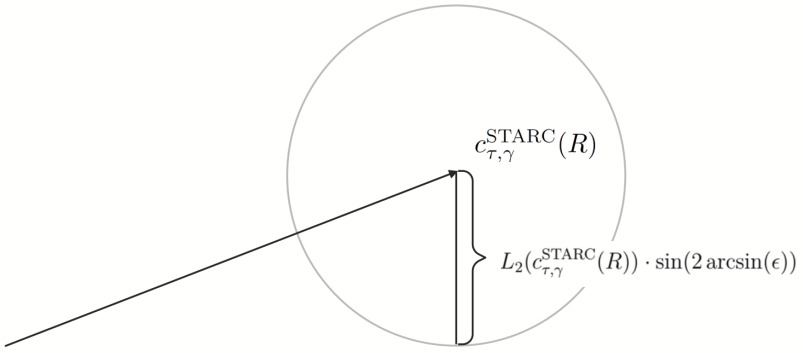
Recall that $d^{\mathrm{STARC}}_{\tau,\gamma}$ is invariant order-preserving transformations, and that all transformations in $S'\mathrm{R}_\tau \bigodot \mathrm{PS}_\gamma \bigodot \mathrm{LS}$ are order-preserving. This means that $d^{\mathrm{STARC}}_{\tau,\gamma}(R, t_i(R)) = 0$ for $i \in \{1, 3\}$.

For t_2, let R be an arbitrary reward function. First note that if R is trivial, then $c^{\mathrm{STARC}}_{\tau,\gamma}(R) = R_0$, and so $L_2(c^{\mathrm{STARC}}_{\tau,\gamma}(R)) = 0$. This implies that $L_2(R, t_2(R)) = 0$, and so $d^{\mathrm{STARC}}_{\tau,\gamma}(R, t_2(R)) = 0$. By the triangle inequality, we then have that $d^{\mathrm{STARC}}_{\tau,\gamma}(R, t(R)) = 0 \leq \epsilon$.

Next, assume that R is non-trivial. Recall that $c^{\mathrm{STARC}}_{\tau,\gamma}$ is a linear orthogonal projection; this means that $L_2(c^{\mathrm{STARC}}_{\tau,\gamma}(R_1), c^{\mathrm{STARC}}_{\tau,\gamma}(R_2)) \leq L_2(R_1, R_2)$. As such, if $L_2(R, t_2(R)) \leq L_2(c^{\mathrm{STARC}}_{\tau,\gamma}(R)) \cdot \sin(2 \arcsin(\epsilon))$, then

$$L_2(c^{\mathrm{STARC}}_{\tau,\gamma}(R), c^{\mathrm{STARC}}_{\tau,\gamma}(t_2(R))) \leq L_2(c^{\mathrm{STARC}}_{\tau,\gamma}(R)) \cdot \sin(2 \arcsin(\epsilon))$$

as well. Consider the set of all reward functions in $\mathrm{Im}(c^{\mathrm{STARC}}_{\tau,\gamma})$ whose distance to $c^{\mathrm{STARC}}_{\tau,\gamma}(R)$ is at most $L_2(c^{\mathrm{STARC}}_{\tau,\gamma}(R)) \cdot \sin(2 \arcsin(\epsilon))$, as in the following diagram:



Now $c^{\mathrm{STARC}}_{\tau,\gamma}(t_2(R))$ is located within the sphere depicted in the diagram above.[2] The vectors R' within this sphere that maximise the distance to $c^{\mathrm{STARC}}_{\tau,\gamma}(R)$ *after normalisation* lie on the tangents of this sphere and the origin. That is, we wish to find an $R' \in \mathrm{Im}(c^{\mathrm{STARC}}_{\tau,\gamma})$ which maximises

$$L_2\left(\frac{R'}{L_2(R')}, \frac{c^{\mathrm{STARC}}_{\tau,\gamma}(R)}{L_2(c^{\mathrm{STARC}}_{\tau,\gamma}(R))}\right),$$

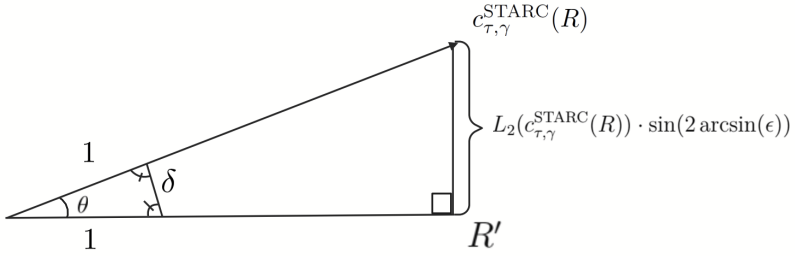
[2]Note that $L_2(c^{\mathrm{STARC}}_{\tau,\gamma}(R)) \cdot \sin(2 \arcsin(\epsilon)) < L_2(c^{\mathrm{STARC}}_{\tau,\gamma}(R))$ for $\epsilon < 0.5$.

subject to the constraint that

$$L_2(c^{\mathrm{STARC}}_{\tau,\gamma}(R), R') \leq L_2(c^{\mathrm{STARC}}_{\tau,\gamma}(R)) \cdot \sin(2\arcsin(\epsilon)).$$

This optimisation problem is solved by the rewards R' such that $L_2(c^{\mathrm{STARC}}_{\tau,\gamma}(R), R') = L_2(c^{\mathrm{STARC}}_{\tau,\gamma}(R)) \cdot \sin(2\arcsin(\epsilon))$, and such that R' forms a right triangle with $c^{\mathrm{STARC}}_{\tau,\gamma}(R)$ and the origin.

Let R' be any such reward, and let δ be the distance between this reward and $c^{\mathrm{STARC}}_{\tau,\gamma}(R)$ after normalisation (i.e., $\delta = L_2\left(R'/L_2(R'), c^{\mathrm{STARC}}_{\tau,\gamma}(R)/L_2(c^{\mathrm{STARC}}_{\tau,\gamma(R))}\right)$). Let θ be the angle between $c^{\mathrm{STARC}}_{\tau,\gamma}(R)$ and R'.



Note that $d^{\mathrm{STARC}}_{\tau,\gamma}(R, R') = 0.5 \cdot \delta$, from the definition of $d^{\mathrm{STARC}}_{\tau,\gamma}$ (also noting that $R' = c(R')$). Elementary trigonometry now tells us that

$$\sin(\theta) = \frac{L_2(c^{\mathrm{STARC}}_{\tau,\gamma}(R)) \cdot \sin(2\arcsin(\epsilon))}{L_2(c^{\mathrm{STARC}}_{\tau,\gamma}(R))},$$

which gives that $\theta = 2\arcsin(\epsilon)$. From this, we have that $\delta = 2\epsilon$, and so $d^{\mathrm{STARC}}_{\tau,\gamma}(R, R') = \epsilon$. Since R' was selected to maximise the $d^{\mathrm{STARC}}_{\tau,\gamma}$-distance to R among all rewards R'' such that $L_2(R, R'') \leq L_2(c^{\mathrm{STARC}}_{\tau,\gamma}(R)) \cdot \sin(2\arcsin(\epsilon))$, we conclude that if $L_2(R, R'') \leq L_2(c^{\mathrm{STARC}}_{\tau,\gamma}(R)) \cdot \sin(2\arcsin(\epsilon))$, then $d^{\mathrm{STARC}}_{\tau,\gamma}(R, R'') \leq \epsilon$. Since R was selected arbitrarily, this proves that $d^{\mathrm{STARC}}_{\tau,\gamma}(R, t_2(R)) \leq \epsilon$ for all R. Since $d^{\mathrm{STARC}}_{\tau,\gamma}(R, t_1(R)) = 0$ and $d^{\mathrm{STARC}}_{\tau,\gamma}(R, t_3(R)) = 0$ for all R, and since $d^{\mathrm{STARC}}_{\tau,\gamma}$ is a pseudometric, we thus have that $d^{\mathrm{STARC}}_{\tau,\gamma}(R, t(R)) \leq \epsilon$ for all R. This completes the other direction, and hence the proof. □

The statement of Theorem 97 is quite terse, so let us briefly unpack it. First of all, $d^{\mathrm{STARC}}_{\tau,\gamma}$ is invariant to any transformation that preserves the policy ordering of the

reward function, and these transformations are exactly those that can be expressed as a combination of potential shaping, S'-redistribution, and positive linear scaling. As such, we can apply an arbitrary number of such transformations. Moreover, we can also transform R in any way that does not change the standardised reward function $s^{\text{STARC}}_{\tau,\gamma}(R)$ by more than ϵ; this is equivalent to the stated condition on t_2. Note that $\sin(2\arcsin(\epsilon)) \approx 2\epsilon$ for small ϵ, so the right-hand side is approximately equal to $2\epsilon \cdot L_2(c^{\text{STARC}}_{\tau,\gamma}(R))$. However, also note that $L_2(c^{\text{STARC}}_{\tau,\gamma}(R)) \leq L_2(R)$. The requirement that $\epsilon < 0.5$ makes the calculation easier, and is included for convenience. Generalising Theorem 97 by removing this requirement would be straightforward, but tedious. However, note that $d^{\text{STARC}}_{\tau,\gamma}$ ranges between 0 and 1, so a $d^{\text{STARC}}_{\tau,\gamma}$-distance greater than 0.5 would be very large.

Using this, we can now state necessary and sufficient conditions that completely characterise all types of misspecification that the Boltzmann-rational model and the MCE model will tolerate:

Corollary 98. *Let $\epsilon < 0.5$, and let T_ϵ be the set of all reward transformations $t : \mathcal{R} \to \mathcal{R}$ that satisfy Theorem 97. Let $f : \mathcal{R} \to \Pi$ be either $b_{\tau,\gamma,\beta}$ or $c_{\tau,\gamma,\alpha}$. Then f is ϵ-robust to misspecification with g (as measured by $d^{\text{STARC}}_{\tau,\gamma}$) if and only if $g = f \circ t$ for some $t \in T_\epsilon$ such that $f \neq g$.*

Proof. Immediate from Lemma 20 and Theorems 97, 73, and 74. □

In principle, Corollary 98 completely describes the misspecification robustness of the Boltzmann-rational model and of the MCE model, as measured by $d^{\text{STARC}}_{\tau,\gamma}$. However, the statement of Corollary 98 is rather opaque, and difficult to interpret qualitatively. For this reason, we will in the subsequent sections examine a few important special types of misspecification, and derive results that are more intuitively intelligible.

We should also briefly comment on the fact that Corollary 98 does not cover $o^\star_{\tau,\gamma}$, i.e. the optimality model. The reason for this is that, unless $|\mathcal{S}| = 1$ and $|\mathcal{A}| = 2$, there are reward functions R_1, R_2 such that $o^\star_{\tau,\gamma}(R_1) = o^\star_{\tau,\gamma}(R_2)$, but

$d^{\text{STARC}}_{\tau,\gamma}(R_1, R_2) > 0$ (Corollary 78). This means that Lemma 20 does not apply to $o^\star_{\tau,\gamma}$ when $d^{\mathcal{R}} = d^{\text{STARC}}_{\tau,\gamma}$. Moreover:

Proposition 99. *Let $d^{\mathcal{R}}$ be a pseudometric on $\mathcal{R}$ that is both sound and complete. Then unless $|\mathcal{S}| = 1$ and $|\mathcal{A}| = 2$, there exists an $E > 0$ such that for all $\epsilon < E$, there is no behavioural model g such that $o^\star_{\tau,\gamma}$ is ϵ-robust to misspecification with g as measured by $d^{\mathcal{R}}$.*

Proof. By Corollary 78, if $|\mathcal{S}| \geq 2$ or $|\mathcal{A}| \geq 3$, and $d^{\mathcal{R}}$ is both sound and complete, then there exists reward functions R_1, R_2 such that $o^\star_{\tau,\gamma}(R_1) = o^\star_{\tau,\gamma}(R_2)$, and such that $d^{\mathcal{R}}(R_1, R_2) > 0$. Thus $d^{\mathcal{R}}(R_1, R_2) = E > 0$, and so $o^\star_{\tau,\gamma}$ violates condition 3 of Definition 8 for all $\epsilon < E$. □

An analogous result will hold for any behavioural model f and any pseudometric $d^{\mathcal{R}}$ for which $f(R_1) = f(R_2) \implies d^{\mathcal{R}}(R_1, R_2) = 0$. Note that E corresponds to the upper diameter of $\text{Am}(o^\star_{\tau,\gamma})$. This means that the exact value of E will depend on the choice of pseudometric $d^{\mathcal{R}}$, and potentially also on the transition function τ, discount γ, and initial state distribution μ_0.

7.2 Perturbation Robustness

It is interesting to know whether or not a behavioural model f is robust to misspecification with any behavioural model g that is "close" to f. But what does it mean for f and g to be "close"? One option is to say that f and g are close if they always produce similar policies. In this section, we will explore under what conditions f is robust to such misspecification, and provide necessary and sufficient conditions. Our results are given relative to a pseudometric d^{Π} on Π. For example, $d^{\Pi}(\pi_1, \pi_2)$ may be the L_2-distance between π_1 and π_2, or it may be the KL divergence between their trajectory distributions, or it may be the L_2-distance between their occupancy measures, and so on. As usual, our results apply for any choice of d^{Π} unless otherwise stated. We can now define a notion of a *perturbation* and a notion of *perturbation robustness*:

Definition 100. Let $f, g : \mathcal{R} \to \Pi$ be two behavioural models, and let d^Π be a pseudometric on Π. Then g is a δ-perturbation of f if $g \neq f$ and for all $R \in \mathcal{R}$ we have that $d^\Pi(f(R), g(R)) \leq \delta$.

Definition 101. Let $f : \mathcal{R} \to \Pi$ be a behavioural model, let $d^\mathcal{R}$ be a pseudometric on $\mathcal{R}$, and let d^Π be a pseudometric on Π. Then f is ϵ-robust to δ-perturbation if f is ϵ-robust to misspecification with g (as measured by $d^\mathcal{R}$) for any behavioural model $g : \mathcal{R} \to \Pi$ that is a δ-perturbation of f (as defined by d^Π) with $\mathrm{Im}(g) \subseteq \mathrm{Im}(f)$.

A δ-perturbation of f is simply any function that is similar to f on all inputs, and f is ϵ-robust to δ-perturbation if a small perturbation of the observed policy leads to a small error in the inferred reward function. It would be desirable for a behavioural model to be robust in this sense. To start with, this captures any form of misspecification that always leads to a small change in the final policy. Moreover, in practice, we can often not observe the exact policy of the demonstrator, and must instead approximate it from a number of samples. In this case, we should also expect to infer a policy that is a perturbation of the true policy. Before moving on, we need one more definition:

Definition 102. Let $f : \mathcal{R} \to \Pi$ be a behavioural model, let $d^\mathcal{R}$ be a pseudometric on $\mathcal{R}$, and let d^Π be a pseudometric on Π. Then f is ϵ/δ-separating if $d^\mathcal{R}(R_1, R_2) > \epsilon \implies d^\Pi(f(R_1), f(R_2)) > \delta$ for all $R_1, R_2 \in \mathcal{R}$.

Intuitively speaking, f is ϵ/δ-separating if reward functions that are far apart, are sent to policies that are far apart.[3] Using this, we can now state our main result for this section:

Theorem 103. *Let $f : \mathcal{R} \to \Pi$ be a behavioural model, let $d^\mathcal{R}$ be a pseudometric on $\mathcal{R}$, and let d^Π be a pseudometric on Π. Then f is ϵ-robust to δ-perturbation (as defined by $d^\mathcal{R}$ and d^Π) if and only if f is ϵ/δ-separating (as defined by $d^\mathcal{R}$ and d^Π).*

[3]Note that this definition is *not* saying that reward functions which are close must be sent to policies which are close. In other words, f being ϵ/δ-separating is *not* a continuity condition. It is also not a local property of f, but rather, a global property. It is, however, a continuity condition on the inverse of f.

Proof. For the first direction, suppose f is ϵ/δ-separating, and let g be a δ-perturbation of f with $\text{Im}(g) \subseteq \text{Im}(f)$. We will show that f and g satisfy the conditions of Definition 8. For the first condition, let R_1, R_2 be two arbitrary reward functions such that $f(R_1) = g(R_2)$. Since g is a δ-perturbation of f, we have that $d^{\Pi}(g(R_2), f(R_2)) \leq \delta$. Since $f(R_1) = g(R_2)$, straightforward substitution thus gives us that $d^{\Pi}(f(R_1), f(R_2)) \leq \delta$. Since f is ϵ/δ-separating, this means that $d^{\mathcal{R}}(R_1, R_2) \leq \epsilon$. Since R_1 and R_2 were chosen arbitrarily, this means that if $f(R_1) = g(R_2)$ then $d^{\mathcal{R}}(R_1, R_2) \leq \epsilon$. Thus, the first condition of Definition 8 holds. For the third condition, note that if $f(R_1) = f(R_2)$, then $d^{\Pi}(f(R_1), f(R_2)) = 0 \leq \delta$. Since f is ϵ/δ-separating, this means that $d^{\mathcal{R}}(R_1, R_2) \leq \epsilon$, which means that the second condition is satisfied as well. The second condition is satisfied, since we assume that $\text{Im}(g) \subseteq \text{Im}(f)$, and the fourth condition is satisfied by the definition of δ-perturbations. This means that f and g satisfy all the conditions of Definition 8, and thus f is ϵ-robust to misspecification with g. Since g was chosen arbitrarily, this means that f is ϵ-robust to misspecification with any δ-perturbation g such that $\text{Im}(g) \subseteq \text{Im}(f)$. Thus f is ϵ-robust to δ-perturbation.

For the second direction, suppose f is *not* ϵ/δ-separating. This means that there exist $R_1, R_2 \in \mathcal{R}$ such that $d^{\mathcal{R}}(R_1, R_2) > \epsilon$ and $d^{\Pi}(f(R_1), f(R_2)) \leq \delta$. Now let $g : \mathcal{R} \to \mathcal{R}$ be the behavioural model where $g(R_1) = f(R_2)$, $g(R_2) = f(R_1)$, and $g(R) = f(R)$ for all $R \notin \{R_1, R_2\}$. Now g is a δ-perturbation of f. However, f is not ϵ-robust to misspecification with g, since $g(R_1) = f(R_2)$, but $d^{\mathcal{R}}(R_1, R_2) > \epsilon$. Thus, if f is not ϵ/δ-separating then f is not ϵ-robust to δ-perturbation, which in turn means that if f is ϵ-robust to δ-perturbation, then f is must be ϵ/δ-separating. □

We have thus obtained necessary and sufficient conditions that describe when a behavioural model is robust to perturbations — namely, it has to be the case that this behavioural model sends reward functions that are far apart, to policies that are far apart. This ought to be quite intuitive; if two policies are close, then perturbations may lead us to conflate them. To be sure that the learnt reward function is close to the true reward function, we therefore need it to be the case that policies that

are close always correspond to reward functions that are close (or, conversely, that reward functions which are far apart correspond to policies which are far apart).

Our next question is, of course, whether or not the standard behavioural models are ϵ/δ-separating. Surprisingly, we will show that this is *not* the case, when the distance between reward functions is measured using $d^{\text{STARC}}_{\tau,\gamma}$, and the policy metric d^{Π} is similar to Euclidean distance. Moreover, this applies to any continuous behavioural model:

Theorem 104. *Let $d^{\mathcal{R}}$ be $d^{\text{STARC}}_{\tau,\gamma}$, and let d^{Π} be a pseudometric on Π which satisfies the condition that for all δ there exists a δ' such that if $L_2(\pi_1, \pi_2) < \delta'$ then $d^{\Pi}(\pi_1, \pi_2) < \delta$. Let $f : \mathcal{R} \to \Pi$ be any continuous behavioural model. Then f is not ϵ/δ-separating for any $\epsilon < 1$ or $\delta > 0$.*

Proof. Let δ be any positive constant. By assumption, there exists a δ' such that if $L_2(\pi_1, \pi_2) < \delta'$ then $d^{\Pi}(\pi_1, \pi_2) < \delta$. Moreover, since f is continuous, there exists an ϵ such that if $L_2(R_1, R_2) < \epsilon$, then $L_2(f(R_1), f(R_2)) < \delta'$. Next, let R be any reward function that is non-trivial under τ and γ. We now have that, for any positive constant c, the reward functions $c \cdot R$ and $-c \cdot R$ have the opposite policy ordering, which means that $d^{\text{STARC}}_{\tau,\gamma}(c \cdot R, -c \cdot R) = 1$. Moreover, by making c sufficiently small, we can ensure that $L_2(c \cdot R, -c \cdot R) < \epsilon$. Thus, for any positive δ there exist reward functions $c \cdot R$ and $-c \cdot R$ such that $d^{\Pi}(f(c \cdot R), f(-c \cdot R)) < \delta$, and such that $d^{\text{STARC}}_{\tau,\gamma}(c \cdot R, -c \cdot R) = 1$. Hence f is not ϵ/δ-separating for any $\delta > 0$ and any $\epsilon < 1$. □

Note that the Boltzmann-rational model and the maximal causal entropy model (i.e. $b_{\tau,\gamma,\beta}$ and $c_{\tau,\gamma,\alpha}$) both are continuous, and hence subject to Theorem 104. The condition given on d^{Π} in Theorem 104 is satisfied by any norm, but will also be satisfied by other metrics.[4]

Intuitively, the fundamental reason for why Theorem 104 holds is that if f is continuous, then it must send reward functions that are close under the L_2-norm

[4]Note that while Theorem 104 uses a "special" pseudometric on $\mathcal{R}$, in the form of $d^{\text{STARC}}_{\tau,\gamma}$, we do not need to use a special (pseudo)metric on Π, because for policies, L_2 does capture the relevant notion of similarity.

to policies that are close under the L_2-norm. However, there are reward functions that are close under the L_2-norm but which have a large STARC distance. Hence f will send some reward functions that are far apart (under $d^{\text{STARC}}_{\tau,\gamma}$) to policies which are close, which means that f is not ϵ/δ-separating. A similar result will hold for any other pseudometric $d^{\mathcal{R}}$ on $\mathcal{R}$ that is both sound and complete, if the upper bound on ϵ is replaced with the smallest distance between any two opposite reward functions under $d^{\mathcal{R}}$. Note that this distance always is 1 under $d^{\text{STARC}}_{\tau,\gamma}$.

It is worth noting that the proof of Theorem 104 only demonstrates that we may run into trouble for reward functions that are very close to R_0, and we may expect such reward functions to be unlikely (both in the sense that the observed agent is unlikely to have such a reward function, and in the sense that the learning algorithm is unlikely to generate such a hypothesis). It would therefore be natural to restrict $\mathcal{R}$ in some way, for example by imposing a minimum size on the L_2-norm of all considered reward functions, or by supposing that they are normalised. We will discuss this option further in Chapter 8, where we also give a generalisation of Theorem 104.

7.3 Misspecified Parameters

In Chapter 6, we showed that many behavioural models are not $\text{OPT}_{\tau,\gamma}$-robust to any misspecification of τ or γ. However, this result says that we cannot identify the exact right optimal policies, or the exact right policy order, given misspecification of τ or γ. This does not rule out the possibility that a small misspecification of τ or γ leads to a small (but nonzero) STARC distance between the true reward function and the learnt reward function. This is the question that we will investigate in this section.

First of all, recall that Lemma 17 implies that if f is ϵ-robust to misspecification with g (as measured by $d^{\mathcal{R}}$), and $g(R_1) = g(R_2)$, then we have that $d^{\mathcal{R}}(R_1, R_2) \leq 2\epsilon$. The converse of this statement is that if there are reward functions R_1, R_2 such that $g(R_1) = g(R_2)$ and $d^{\mathcal{R}}(R_1, R_2) > 2\epsilon$, then f is *not* ϵ-robust to misspecification with g (as measured by $d^{\mathcal{R}}$). Therefore, we can use the (upper) diameter of $\text{Am}(g)$ to derive

a limit on how robust any f may be to misspecification with g. Our results in this section will use this proof strategy. We first consider the case when τ is misspecified:

Theorem 105. *If $f_\tau : \mathcal{R} \to X$ is invariant to S'-redistribution with τ, and $\tau_1 \neq \tau_2$, then f_{τ_1} is not ϵ-robust to misspecification with f_{τ_2} under $d^{\mathrm{STARC}}_{\tau_3,\gamma}$ for any τ_3, any γ, and any $\epsilon < 0.5$.*

Proof. If $\tau_1 \neq \tau_2$, then either $\tau_1 \neq \tau_3$ or $\tau_2 \neq \tau_3$. If $\tau_1 \neq \tau_3$, then Theorem 84 implies that there for each $\delta > 0$ exists reward functions R_1, R_2 such that $f_{\tau_1}(R_1) = f_{\tau_1}(R_2)$, but such that $d^{\mathrm{STARC}}_{\tau_3,\gamma}(R_1, R_2) \geq 1 - \delta$. Thus f_{τ_1} violates condition 2 of Definition 8 for all $\epsilon < 1$. Similarly, if $\tau_2 \neq \tau_3$, then Theorem 84 implies that there for each $\delta > 0$ exists reward functions R_1, R_2 such that $f_{\tau_2}(R_1) = f_{\tau_2}(R_2)$, but such that $d^{\mathrm{STARC}}_{\tau_3,\gamma}(R_1, R_2) \geq 1 - \delta$. Then Lemma 17 implies that there can be no f that is ϵ-robust to misspecification with f_{τ_2} (as defined by $d^{\mathrm{STARC}}_{\tau_3,\gamma}$) for any $\epsilon < 0.5$. □

Theorem 105 is saying that if some behavioural model f is invariant to S'-redistribution, then it is not robust to any degree of misspecification of τ (even if τ_1 and τ_2 are arbitrarily close). Note that a $d^{\mathrm{STARC}}_{\tau,\gamma}$-distance of 0.5 is very large; this corresponds to the case where the reward functions are nearly orthogonal. The greatest possible value of $d^{\mathrm{STARC}}_{\tau,\gamma}$ is 1. Moreover, optimal policies, Boltzmann-rational policies, and maximal causal entropy policies, are all invariant to S'-redistribution, and hence $o^\star_{\tau,\gamma}$, $b_{\tau,\gamma,\beta}$, and $c_{\tau,\gamma,\alpha}$ are subject to Theorem 105. This means that Theorem 94 generalises to the setting with distance metrics. Moreover, contrary to what we might expect, a small amount of misspecification of τ does not guarantee a small error in the learnt reward R_H, if this error is quantified with STARC metrics.

We next consider the case when the discount parameter, γ, is misspecified. As before, we say that a transition function τ is *trivial* if for all states s and all actions a_1, a_2, we have that $\tau(s, a_1) = \tau(s, a_2)$.

Theorem 106. *If $f_\gamma : \mathcal{R} \to \Pi$ is invariant to potential shaping with γ, and $\gamma_1 \neq \gamma_2$, then f_{γ_1} is not ϵ-robust to misspecification with f_{γ_2} under $d^{\mathrm{STARC}}_{\tau,\gamma_3}$ for any non-trivial τ, any γ_3, and any $\epsilon < 0.5$.*

Proof. If $\gamma_1 \neq \gamma_2$, then either $\gamma_1 \neq \gamma_3$ or $\gamma_2 \neq \gamma_3$. If $\gamma_1 \neq \gamma_3$, then Theorem 87 implies that there for each $\delta > 0$ exists reward functions R_1, R_2 such that $f_{\gamma_1}(R_1) = f_{\gamma_1}(R_2)$, but such that $d^{\text{STARC}}_{\tau,\gamma_3}(R_1, R_2) \geq 1 - \delta$. Thus f_{γ_1} violates condition 2 of Definition 8 for all $\epsilon < 1$. Similarly, if $\gamma_2 \neq \gamma_3$, then Theorem 87 implies that there for each $\delta > 0$ exists reward functions R_1, R_2 such that $f_{\gamma_2}(R_1) = f_{\gamma_2}(R_2)$, but such that $d^{\text{STARC}}_{\tau,\gamma_3}(R_1, R_2) \geq 1 - \delta$. Then Lemma 17 implies that there can be no f that is ϵ-robust to misspecification with f_{γ_2} (as defined by $d^{\text{STARC}}_{\tau,\gamma_3}$) for any $\epsilon < 0.5$. □

Of course, any interesting environment will have a non-trivial transition function, so this requirement is very mild. This means that Theorem 106 is saying that if a behavioural model f is invariant to potential shaping, then it is not robust to any misspecification of the discount parameter. Note that this holds even if γ_1 and γ_2 are arbitrarily close! Moreover, optimal policies, Boltzmann-rational policies, and MCE policies are all invariant to potential shaping, and hence $o_{\tau,\gamma}$, $b_{\tau,\gamma,\beta}$, and $c_{\tau,\gamma,\alpha}$ are subject to Theorem 106. This means that Theorem 95 generalises to the setting with distance metrics. Moreover, contrary to what we might expect, a small amount of misspecification of γ does not guarantee a small error in the learnt reward R_H, if this error is quantified with STARC metrics.

Before moving on, let us briefly give some additional intuition for why Theorems 105 and 106 are true. Roughly speaking, the core reason for why Theorem 105 is true is that, if $\tau_1 \neq \tau_2$, then there are reward functions which are equivalent under τ_2, but very different under τ_1. We give an intuitive example of such a case in Section 5.4. More formally, for each reward R_1 we can find a reward R_2 such that R_1 and R_2 differ by S'-redistribution under τ_2, but such that $d^{\text{STARC}}_{\tau_1,\gamma}(R_1, R_2)$ is arbitrarily close to 1. Now suppose the observed policy is computed by a behavioural model g_{τ_2} that is invariant to S'-redistribution with τ_2, and that the learning algorithm $\mathcal{L}$ is given training data from a policy π such that $\pi = g_{\tau_2}(R_1) = g_{\tau_2}(R_2)$. What reward function R_H should $\mathcal{L}$ converge to? It is impossible to find any R_H that is close to both R_1 and R_2 under $d^{\text{STARC}}_{\tau_1,\gamma}$, and so $\mathcal{L}$ must necessarily be unable to guarantee that the learnt reward function is close to the true reward function. Similarly, the reason for why Theorem 106 is true is that, if $\gamma_1 \neq \gamma_2$ (and τ is nontrivial), then

there are reward functions which are equivalent when discounting with γ_2, but very different when discounting with γ_1 (we also give an intuitive example of such a case in Section 5.4). Thus, a misspecified γ causes similar problems as a misspecified τ.

7.4 Transfer Learning

STARC metrics, such as $d^{\text{STARC}}_{\tau,\gamma}$, are designed to be sound and complete. Moreover, our definitions of soundness and completeness for a pseudometric $d^{\mathcal{R}}$ require that $d^{\mathcal{R}}(R_1, R_2)$ is small if and only if the regret of using R_1 instead of R_2 is small, relative to a particular transition function τ and discount factor γ. A natural question is what happens if we strengthen this requirement, and demand that the regret is small for any choice of τ or any choice of γ. We briefly discuss this setting here.

In short, as for Definition 7, it is impossible to guarantee transfer to any τ or γ. This is already implied by the results in Section 5.4. In particular, if $f_{\tau,\gamma}$ is invariant to S'-redistribution (with τ) and potential shaping (with γ), then the (upper and lower) diameter of $\text{Am}(f_{\tau_1,\gamma_1})$ under $d^{\text{STARC}}_{\tau_2,\gamma_2}$ is 1, provided that either $\tau_1 \neq \tau_2$, or $\gamma_1 \neq \gamma_2$ and τ_2 is non-trivial. Then f_{τ_1,γ_1} will violate condition 3 in Definition 8. Moreover, note that this result is not specific to $d^{\text{STARC}}_{\tau,\gamma}$, and that a similar result will hold for any pseudometric on $\mathcal{R}$ that is both sound and complete.

All models are wrong, but some are useful.

— George Box, 1979.

8 Generalising Our Analysis

In this chapter, we discuss how to generalise our analysis, and extend the frameworks presented in Chapter 3. In particular, the definitions that we have worked with so far quantify over all reward functions, both for the true reward function $R^\star$ and the learnt reward function R_H. In some cases, we may have some prior knowledge about the true reward function $R^\star$, or we may know that the inductive bias of the learning algorithm is unlikely to generate certain reward functions R_H, even if they are compatible with the training data. Consequently, we may wish to incorporate assumptions about the true reward or about the inductive bias of the learning algorithm into our analysis. We will discuss these extensions, and show that our analysis remains largely unchanged by such generalisations. We will also discuss some alternative equivalence relations on $\mathcal{R}$.

8.1 Assumptions About Inductive Bias

It is worth noting that none of the definitions in Chapter 3 make any assumptions about the *inductive bias* of the learning algorithm. By "inductive bias", we here refer to everything that determines which hypothesis a learning algorithm selects, when there are several hypotheses in its hypothesis space that are compatible with a given learning objective and set of training data. In particular, in the case of

IRL, there will typically be multiple reward functions that fit a given set of training data under a given behavioural model – all parts of the learning algorithm that determine which of these reward functions is learnt together make up the inductive bias of the learning algorithm.[1] For example, let the true reward function be $R^\star$, let the true data generating process be described by g, and let f be the assumed model of the data generating process. Then both Definition 7 and 8 require that *every* reward function R_H that is compatible with the training data must be equivalent or similar to the true reward. This requirement may seem unnecessarily strong, because some reward functions R_H such that $f(R_H) = g(R^\star)$ may be very unlikely to be generated by the learning algorithm, $\mathcal{L}$. This, in turn, raises the question of whether we may be able to create weaker, more permissive formalisations of ambiguity tolerance and misspecification robustness by also making assumptions about the inductive bias of the learning algorithm (i.e., assumptions about how the learning algorithm selects a reward function when multiple reward functions fit the training data). In this section, we discuss this option.

Let us first focus on Definition 3, which formalises when a given application tolerates the ambiguity of a given data source. In this case, it does not seem like anything can be gained from incorporating assumptions about inductive bias. To see this, consider the following modified definition:

Definition 107. Given a reward object $f : \mathcal{R} \to X$, we say that $I : X \to \mathcal{R}$ is an *inductive bias* for f if $f(I(x)) = x$ for all $x \in X$.

Definition 108. Given two reward objects $f : \mathcal{R} \to X$, $g : \mathcal{R} \to Y$, and an *inductive bias* $I : X \to \mathcal{R}$ for f, we say that g tolerates the ambiguity of f with I if, for all $R^\star \in \mathcal{R}$, if $R_H = I(f(R^\star))$, then $g(R_H) = g(R^\star)$.

[1] Note that different texts in the machine learning literature may consider different things to be part of the "inductive bias" of a learning algorithm. For example, some authors may consider the behavioural model to be a part of the inductive bias of an IRL algorithm. In this text, we instead consider the behavioural model to be a part of the learning objective. That is, the learning objective specifies when a hypothesis (which in our case is a reward function) is considered to fit a set of training data (which in our case is a policy), and the inductive bias is that which determines which hypothesis the learning algorithm selects. However, this is just a terminological choice, and it has no impact on the mathematical results we derive.

Note that $I : X \to \mathcal{R}$ is an inductive bias for $f : \mathcal{R} \to X$ if, for any $x \in X$, I maps x to a reward function R such that $f(R) = x$. In other words, I is a function that, for any possible observable data x, picks a reward function R that is compatible with x under f. Definition 108 then says that g tolerates the ambiguity of f with I if, for any true reward function $R^\star$ and any corresponding data distribution $f(R^\star)$, I always picks a reward function R_H such that $g(R_H) = g(R^\star)$. This is directly analogous to Definition 3, except that we assume that the learning algorithm uses a particular inductive bias I. Using these definitions, we can now derive the following result:

Theorem 109. *Let $f : \mathcal{R} \to X$, $g : \mathcal{R} \to Y$ be any two reward objects, and let $I : X \to \mathcal{R}$ be any inductive bias for f. Then g tolerates the ambiguity of f with I (in the sense of Definition 108) if and only if $\mathrm{Am}(f) \preceq \mathrm{Am}(g)$ (in the sense of Definition 3).*

Proof. For the first direction, assume that $\mathrm{Am}(f) \preceq \mathrm{Am}(g)$. Let $R^\star$ be an arbitrary reward function, and let R_H be the reward such that $I(f(R^\star)) = R_H$. Note that $f(I(x)) = x$ for all $x \in X$, which means that $f(I(f(R^\star))) = f(R^\star)$. In other words, $f(R_H) = f(R^\star)$. Since $\mathrm{Am}(f) \preceq \mathrm{Am}(g)$, this means that $g(R_H) = g(R^\star)$. This completes the first direction.

For the other direction, assume that g tolerates the ambiguity of f with I, in the sense of Definition 108. Suppose $f(R_1) = f(R_2)$. Since $g(I(f(R))) = g(R)$ for all R, we have that $g(I(f(R_1))) = g(R_1)$ and $g(I(f(R_2))) = g(R_2)$. Moreover, since $f(R_1) = f(R_2)$, this means that $g(I(f(R_1))) = g(I(f(R_2)))$. By transitivity, this then implies that $g(R_1) = g(R_2)$, and so $\mathrm{Am}(f) \preceq \mathrm{Am}(g)$. □

Thus, Definition 108 is functionally equivalent to Definition 3. In other words, for the purposes of ambiguity tolerance, it does not make any difference which inductive bias I the learning algorithm uses. Also note that, while Definition 107 defines I to be a function that deterministically picks a fixed R for each $x \in X$, we would obtain a result analogous to Theorem 109 if we instead defined I to be a set-valued function, etc. Thus, the analysis in Chapter 5 which is based on

Definition 3 would remain unchanged if we defined ambiguity tolerance relative to a particular choice of inductive bias for the learning algorithm.

Let us next consider Definition 7, which defines misspecification robustness relative to equivalence relations on $\mathcal{R}$. In this case, we similarly find that the inductive bias does not affect our results. To see this, consider the following modified definition of misspecification robustness:

Definition 110. Given a partition P of $\mathcal{R}$, two reward objects $f, g : \mathcal{R} \to X$, and an *inductive bias* $I : X \to \mathcal{R}$ for f, we say that f is *P-robust to misspecification* with g using I if each of the following conditions are satisfied:

1. $I(g(R)) \equiv_P R$ for all R.
2. $\text{Im}(g) \subseteq \text{Im}(f)$.
3. $I(f(R)) \equiv_P R$ for all R.
4. $f \neq g$.

Definition 110 is simply directly analogous to Definition 7, except that it makes the assumption that the learning algorithm $\mathcal{L}$ uses the inductive bias described by I. Using this definitions, we then get the following result:

Theorem 111. *Let P be any partition of $\mathcal{R}$, let $f, g : \mathcal{R} \to X$ be any two reward objects, and let $I : X \to \mathcal{R}$ be any inductive bias for f. Then f is P-robust to misspecification with g (in the sense of Definition 7) if and only if f is* P-robust to misspecification *with g using I (in the sense of Definition 110).*

Proof. For the first direction, assume that f is P-robust to misspecification in the sense of Definition 7. We then have that:

1. If $f(R_1) = g(R_2)$ then $R_1 \equiv_P R_2$.
2. $\text{Im}(g) \subseteq \text{Im}(f)$.
3. $\text{Am}(f) \preceq P$.

4. $f \neq g$.

To show that f is P-robust to misspecification with g using I, we must show that the following two conditions hold:

1. $I(g(R)) \equiv_P R$ for all R.

2. $I(f(R)) \equiv_P R$ for all R.

Let R be an arbitrary reward function. Note that $f(I(x)) = x$ for all $x \in X$. This means that $I(g(R))$ is a reward function such that $f(I(g(R))) = g(R)$. Since $R_1 \equiv_P R_2$ whenever $f(R_1) = g(R_2)$, this means that $I(g(R)) \equiv_P R$. Thus, the first condition holds. Similarly, $I(f(R))$ is a reward function such that $f(I(f(R))) = f(R)$. Since $\mathrm{Am}(f) \preceq P$, this means that $I(f(R)) \equiv_P R$, and so the second condition also holds. This completes the first direction.

For the other direction, assume that f is P-robust to misspecification with g using I in the sense of Definition 110. We then have that

1. $I(g(R)) \equiv_P R$ for all R.

2. $\mathrm{Im}(g) \subseteq \mathrm{Im}(f)$.

3. $I(f(R)) \equiv_P R$ for all R.

4. $f \neq g$.

To show that f is P-robust to misspecification with g, we must show that the following two conditions hold:

1. If $f(R_1) = g(R_2)$ then $R_1 \equiv_P R_2$.

2. $\mathrm{Am}(f) \preceq P$.

First, suppose $f(R_1) = f(R_2)$. Since $I(f(R)) \equiv_P R$ for all R, we have that $I(f(R_1)) \equiv_P R_1$ and $I(f(R_2)) \equiv_P R_2$. Moreover, since $f(R_1) = f(R_2)$, this means that $I(f(R_1)) = I(f(R_2))$. By transitivity, this then implies that $R_1 \equiv_P R_2$, and so $\mathrm{Am}(f) \preceq P$. Similarly, suppose $f(R_1) = g(R_2)$. Since $I(g(R)) \equiv_P R$ for all R,

we have that $I(g(R_2)) \equiv_P R_2$. Moreover, since $f(I(f(R_1))) = f(R_1)$, and since $\text{Am}(f) \preceq P$, we have that $I(f(R_1)) \equiv_P R_1$. Since $f(R_1) = g(R_2)$, we thus have that $R_1 \equiv_P R_2$. This completes the proof. □

In other words, our analysis of misspecification robustness in terms of equivalence relations on $\mathcal{R}$ is also not affected by the inductive bias of the learning algorithm. As such, all of our results in Chapter 6 would be identical if we used Definition 110 instead of Definition 7.

The case is less straightforward when we characterise the difference between reward functions in terms of pseudometrics on $\mathcal{R}$, rather than equivalence relations on $\mathcal{R}$ (as with Definition 6 and Definition 8). For example, we can have a reward object $f : \mathcal{R} \to X$ for which the lower diameter of $\text{Am}(f)$ is greater than ϵ, but where there exists an inductive bias I for f such that $d^{\mathcal{R}}(I(f(R)), R) \leq \epsilon$ for all R. To see this, suppose the (upper and lower) diameter of $\text{Am}(f)$ is 2ϵ, but that the inductive bias I always picks a reward function that is in the "middle" of each set in $\text{Am}(f)$, such that the distance between this reward function and all other reward functions in the same set of $\text{Am}(f)$ always is at most ϵ. In this way, the worst-case error between the learnt reward function R_H and the true reward function $R^\star$ may be more than ϵ for data generated under f, even though it can be guaranteed to be at most ϵ under a particular inductive bias I. In a similar way, the conditions for when f is ϵ-robust to misspecification with g (under Definition 8) may also be affected by the inductive bias I of the learning algorithm.

However, note that this generalisation cannot affect the derived results by a substantial amount. To see this, first note that if the upper diameter of $\text{Am}(f)$ under $d^{\mathcal{R}}$ is δ, then for any inductive bias I for f, there is a reward function R such that $d^{\mathcal{R}}(R, I(f(R))) \geq \delta/2$, by the triangle inequality. This means that we will still have to require that the upper diameter of $\text{Am}(f)$ is small. Furthermore, if there are reward functions R_1, R_2 such that $f(R_1) = g(R_2)$ and $d^{\mathcal{R}}(R_1, R_2) = \epsilon$, and if the upper diameter of $\text{Am}(f)$ is δ, then for any inductive bias I for f, we

have that $d^{\mathcal{R}}(R_2, I(g(R_2))) \geq \epsilon - \delta$, again by the triangle inequality.[2] Thus, if the upper diameter of $\text{Am}(f)$ is small, then the inductive bias cannot have a large impact on the misspecification robustness of the algorithm. In other words, we must require the upper diameter of $\text{Am}(f)$ to be small, but if this is the case, then the inductive bias cannot matter much.

More generally, the inductive bias of the learning algorithm could make a meaningful difference to the derived results primarily when the (upper) diameters of $\text{Am}(f)$ and $\text{Am}(g)$ are small, but greater than zero. However, in most of the cases we have analysed, the diameter of $\text{Am}(f)$ is either zero, or very large. For example, the upper diameter of both $\text{Am}(b_{\tau,\gamma,\beta})$ and $\text{Am}(c_{\tau,\gamma,\alpha})$ is zero (Corollary 76), and the upper diameter of $\text{Am}(o^{\star}_{\tau,\gamma})$ is large (Corollary 78). Similarly, when f is invariant to potential shaping with γ or S'-redistribution with τ, and γ or τ is misspecified, then the upper diameter of $\text{Am}(f)$ is large (Theorem 84 and 87). Therefore, we should not expect any of the results we have derived using Definition 8 to change substantially if we modify Definition 8 to also incorporate assumptions about the inductive bias of the learning algorithm. Nonetheless, carrying out this analysis in more detail may be an interesting direction for future work.

8.2 Assumptions About the True Reward

It is worth noting that our definitions in Chapter 3 make no assumptions about the true reward function, $R^{\star}$. While this is reasonable, it does raise the worry that some of our negative results (e.g., those in Sections 5.4, 6.3, and 7.3) may be caused by specific edge-cases that are unlikely to arise in practice. For example, in order for f to be P-robust to misspecification with g, it is required that there for *every* reward function $R^{\star}$ is no R_H such that $f(R_H) = g(R^{\star})$, but $R_H \not\equiv_P R^{\star}$. This may seem unnecessarily strong, if we have reason to believe that certain reward functions $R^{\star}$ are unlikely to come up in practice. A natural question is therefore

[2]For clarity, let us spell this out. First, note that if $g(R_2) = f(R_1)$, and I is an inductive bias for f, then $I(g(R_2))$ is some reward function R_3 such that $f(R_1) = f(R_3)$. Since the upper diameter of $\text{Am}(f)$ is δ, we have that $d^{\mathcal{R}}(R_1, R_3) \leq \delta$. We also have that $d^{\mathcal{R}}(R_1, R_2) = \epsilon$. By the triange inequality, $d^{\mathcal{R}}(R_1, R_2) \leq d^{\mathcal{R}}(R_1, R_3) + d^{\mathcal{R}}(R_3, R_2)$, so $\epsilon \leq x + d^{\mathcal{R}}(R_3, R_2)$, where $x \leq \delta$. This implies that $\epsilon - \delta \leq d^{\mathcal{R}}(R_2, R_3)$.

if we may be able to obtain stronger results by incorporating some assumptions about $R^\star$. In this section, we will discuss this question.

In short, most of our negative results would not change if we incorporate assumptions about the true reward function $R^\star$. This is largely ensured by the fact that we distinguish between the upper and lower diameter of $\mathrm{Am}(f)$. For example, let us first consider the results in Section 5.4, which show that a wide range of behavioural models are unable to guarantee robust transfer to a different transition function τ or discount factor γ. These results are not merely saying that there exists *some* $R^\star$ and *some* R_H such that $R^\star$ and R_H are indistinguishable by the learning algorithm, but such that $R^\star$ and R_H are qualitatively different after transfer to a different τ or γ. Rather, they are saying that there for *every* $R^\star$ is an R_H such that $R^\star$ and R_H are indistinguishable by the learning algorithm, but such that $R^\star$ and R_H are qualitatively different after transfer. In other words, it is not just the *upper* diameter of $\mathrm{Am}(f)$ that is too large, but also the *lower* diameter. Every reward function $R^\star$ is indistinguishable from some reward R_H such that $R^\star$ and R_H are too different to guarantee robust transfer.

The negative results in Section 6.3 and 7.3, which concern misspecified τ or γ, will generalise for a similar reason. Recall that these results are saying, roughly, that if $f_{\tau,\gamma}$ is invariant to S'-redistribution with τ or potential shaping with γ, and either τ or γ is misspecified, then f_{τ_1,γ_1} is not robust to misspecification with f_{τ_2,γ_2}. Intuitively speaking, the reason for why this is true is that at least one of $\mathrm{Am}(f_{\tau_1,\gamma_1})$ or $\mathrm{Am}(f_{\tau_2,\gamma_2})$ will be too large. This result is derived from the ambiguity results in Section 5.4, which means that it is not just the worst-case size (i.e. the upper diameter) of $\mathrm{Am}(f_{\tau_1,\gamma_1})$ or $\mathrm{Am}(f_{\tau_2,\gamma_2})$ that is too large, but the best-case size (i.e. the lower diameter) as well. These results will therefore also not be affected by any assumptions we could make about $R^\star$.

The only exception to this is Theorem 104, which says that no continuous behavioural model is ϵ/δ-separating relative to $d^{\mathrm{STARC}}_{\tau,\gamma}$, which also means that no continuous behavioural model is perturbation robust relative to $d^{\mathrm{STARC}}_{\tau,\gamma}$. The proof of Theorem 104 finds a specific counterexample in the vicinity of the zero

reward, R_0, which could be ruled out by making certain assumptions about $R^\star$. This issue will also be discussed in Section 8.3.

8.3 Restricting the Space of Reward Functions

In Section 8.1, we discussed the option of incorporating assumptions about the inductive bias of the learning algorithm, and in Section 8.2, we discussed the option of incorporating assumptions about the true reward function. Moreover, we argued that neither of these generalisations would make a meaningful difference to our results. But what if we do both at the same time? Formally, suppose that instead of quantifying over all reward functions in $\mathcal{R}$, we restrict the reward functions to lie in some set $\hat{\mathcal{R}} \subseteq \mathcal{R}$. We can then assume that the true reward function $R^\star$ is in $\hat{\mathcal{R}}$, and that the learning algorithm $\mathcal{L}$ only will generate reward functions R_H that are also in $\hat{\mathcal{R}}$. In this section, we will discuss this option. As we will see, our results are largely unaffected by this generalisation, though it does open up some avenues for further analysis.

We first need to generalise the definitions in Chapter 3, which is straightforward; simply replace any quantifier which ranges over all of $\mathcal{R}$ with one that only ranges over $\hat{\mathcal{R}}$ (where $\hat{\mathcal{R}}$ is permitted to be any subset of $\mathcal{R}$). We do this as follows:

Definition 112. Given two partitions P, Q of $\mathcal{R}$, and a set $\hat{\mathcal{R}} \subseteq \mathcal{R}$, we say that $P \preceq Q$ on $\hat{\mathcal{R}}$ if $R_1 \equiv_P R_2$ implies that $R_1 \equiv_Q R_2$ for all $R_1, R_2 \in \hat{\mathcal{R}}$.

Definition 113. Given a pseudometric $d^{\mathcal{R}}$ on $\mathcal{R}$, a set $\hat{\mathcal{R}} \subseteq \mathcal{R}$, and a reward object $f : \mathcal{R} \to X$, we say that the *upper diameter* of $\mathrm{Am}(f)$ on $\hat{\mathcal{R}}$ is

$$\sup\{\mathrm{diam}(S \cap \hat{\mathcal{R}}) : S \in \mathrm{Am}(f), S \cap \hat{\mathcal{R}} \neq \varnothing\}.$$

Similarly, the *lower diameter* of $\mathrm{Am}(f)$ on $\hat{\mathcal{R}}$ is

$$\inf\{\mathrm{diam}(S \cap \hat{\mathcal{R}}) : S \in \mathrm{Am}(f), S \cap \hat{\mathcal{R}} \neq \varnothing\}.$$

Here diam is defined as in Definition 6. If $\hat{\mathcal{R}} = \varnothing$, then the upper and lower diameter of $\mathrm{Am}(f)$ on $\hat{\mathcal{R}}$ is undefined.

Definition 114. Given two reward objects $f, g : \mathcal{R} \to X$, a set $\hat{\mathcal{R}} \subseteq \mathcal{R}$, and a partition P of $\mathcal{R}$, we say that f is P-robust to misspecification with g on $\hat{\mathcal{R}}$ if

1. $f(R_1) = g(R_2)$ implies that $R_1 \equiv_P R_2$ for all $R_1, R_2 \in \hat{\mathcal{R}}$,
2. $\text{Im}(g|_{\hat{\mathcal{R}}}) \subseteq \text{Im}(f|_{\hat{\mathcal{R}}})$,
3. $\text{Am}(f) \preceq P$ on $\hat{\mathcal{R}}$, and
4. $f|_{\hat{\mathcal{R}}} \neq g|_{\hat{\mathcal{R}}}$.

Definition 115. Given two reward objects $f, g : \mathcal{R} \to X$, a set $\hat{\mathcal{R}} \subseteq \mathcal{R}$, and a pseudometric $d^{\mathcal{R}}$ on $\mathcal{R}$, we say that f is ϵ-robust to misspecification with g on $\hat{\mathcal{R}}$ (as measured by $d^{\mathcal{R}}$) if

1. $f(R_1) = g(R_2)$ implies that $d^R(R_1, R_2) \leq \epsilon$ for all $R_1, R_2 \in \hat{\mathcal{R}}$,
2. $\text{Im}(g|_{\hat{\mathcal{R}}}) \subseteq \text{Im}(f|_{\hat{\mathcal{R}}})$,
3. $f(R_1) = f(R_2)$ implies that $d^R(R_1, R_2) \leq \epsilon$ for all $R_1, R_2 \in \hat{\mathcal{R}}$,
4. $f|_{\hat{\mathcal{R}}} \neq g|_{\hat{\mathcal{R}}}$.

Using these more general definitions, we can now generalise our analysis. We should first note that all lemmas in Chapter 3 apply with these more general definitions for any arbitrary subset $\hat{\mathcal{R}} \subseteq \mathcal{R}$. We will not spell out these proofs explicitly, but they are directly analogous to the proofs given in Chapter 3, since none of these proofs rely on any assumptions about $\mathcal{R}$.[3] In addition to this, most of our results in Chapters 5, 6, and 7 carry over to this setting very directly, with only minor modifications. We can first make two straightforward observations:

Proposition 116. *Let P, Q be any partitions on $\mathcal{R}$, and let $\hat{\mathcal{R}}$ be any subset of $\mathcal{R}$. We then have that if $P \preceq Q$ on $\mathcal{R}$, then $P \preceq Q$ on $\hat{\mathcal{R}}$.*

[3]We should also note that the propositions given in Section 3.3 (in particular, Proposition 16, 18, and 19) do not hold for arbitrary sets $\hat{\mathcal{R}}$. However, none of our later results rely on these propositions, since their purpose primarily is to highlight some of the differences between Definition 7 and 8.

Proof. Immediate from the definitions. □

Proposition 117. *For any reward object $f : \mathcal{R} \to X$, any pseudometric $d^{\mathcal{R}}$ $\mathcal{R}$, and any nonempty $\hat{\mathcal{R}} \subseteq \mathcal{R}$, the upper diameter of* $\mathrm{Am}(f)$ *on $\hat{\mathcal{R}}$ is no greater than the upper diameter of* $\mathrm{Am}(f)$ *on $\mathcal{R}$, and likewise for the lower diameter.*

Proof. Immediate from the definitions. □

Proposition 116 says that if g tolerates the ambiguity of f relative to the space of all reward functions $\mathcal{R}$ (in the sense that $\mathrm{Am}(f) \preceq \mathrm{Am}(g)$), then this is (of course) also the case for any restricted space of reward functions $\hat{\mathcal{R}}$. In other words, ambiguity tolerance cannot decrease if the space of all reward functions is restricted. Similarly, Proposition 117 says that the ambiguity of f cannot increase if the space of reward functions is restricted. Intuitively speaking, this means that all our positive results about ambiguity and ambiguity tolerance generalise to arbitrary restrictions on the space of considered reward functions.

However, the converse does *not* hold; it is possible that $\mathrm{Am}(f) \preceq \mathrm{Am}(g)$ on $\hat{\mathcal{R}}$ for some $\hat{\mathcal{R}} \subseteq \mathcal{R}$, even though $\mathrm{Am}(f) \not\preceq \mathrm{Am}(g)$ on $\mathcal{R}$. This is easy to see: $\mathrm{Am}(f) \not\preceq \mathrm{Am}(g)$ on $\mathcal{R}$ if there are reward functions $R_1, R_2 \in \mathcal{R}$ such that $f(R_1) = f(R_2)$ but $g(R_1) \neq g(R_2)$. It may be that $\mathcal{R}$ contains such a counterexample, even though $\hat{\mathcal{R}}$ does not. In the same way, it may be that the upper diameter of $\mathrm{Am}(f)$ on $\hat{\mathcal{R}}$ is ϵ, even though the upper diameter of $\mathrm{Am}(f)$ on $\mathcal{R}$ is greater than ϵ. This means that a *negative* result about ambiguity or ambiguity tolerance may not generalise to a given restricted space of reward functions.

We next show how to extend our results about misspecification to the setting where the space of rewards is restricted. This requires a bit more work:

Theorem 118. *For any $\hat{\mathcal{R}} \subseteq \mathcal{R}$ and any partition P of $\mathcal{R}$, if f is P-robust to misspecification with g on $\mathcal{R}$ then f is P-robust to misspecification with g on $\hat{\mathcal{R}}$, unless $f|_{\hat{\mathcal{R}}} = g|_{\hat{\mathcal{R}}}$. Similarly, if f is P-robust to misspecification with g on $\hat{\mathcal{R}}$ then f is P-robust to misspecification with g' on $\mathcal{R}$ for some g' where $g'|_{\hat{\mathcal{R}}} = g|_{\hat{\mathcal{R}}}$, unless $\mathrm{Am}(f) \not\preceq P$ on $\mathcal{R}$.*

Proof. For the first claim, suppose that f is P-robust to misspecification with g on $\mathcal{R}$, and that $f|_{\hat{\mathcal{R}}} \neq g|_{\hat{\mathcal{R}}}$. Since f is P-robust to misspecification with g on $\mathcal{R}$, we have that for all $R_1, R_2 \in \mathcal{R}$, if $f(R_1) = g(R_2)$ then $R_1 \equiv_P R_2$. Since $\hat{\mathcal{R}} \subseteq \mathcal{R}$, this directly implies that for all $R_1, R_2 \in \hat{\mathcal{R}}$, if $f(R_1) = g(R_2)$ then $R_1 \equiv_P R_2$. We also have that $\text{Im}(f) \subseteq \text{Im(g)}$, which directly implies that $\text{Im}(g|_{\hat{\mathcal{R}}}) \subseteq \text{Im}(f|_{\hat{\mathcal{R}}})$. Moreover, we have that $\text{Am}(f) \preceq P$ on $\mathcal{R}$, which (as shown in Proposition 116) implies that $\text{Am}(f) \preceq P$ on $\hat{\mathcal{R}}$. We assume that $f|_{\hat{\mathcal{R}}} \neq g|_{\hat{\mathcal{R}}}$. Thus, f is P-robust to misspecification with g on $\hat{\mathcal{R}}$.

For the second claim, suppose f is P-robust to misspecification with g on $\hat{\mathcal{R}}$, and that $\text{Am}(f) \preceq P$ on $\mathcal{R}$. We construct a g' as follows; let $g'(R) = g(R)$ for all $R \in \hat{\mathcal{R}}$, and let $g'(R) = f(R)$ for all $R \notin \hat{\mathcal{R}}$. By construction, we have that $g(R) = g'(R)$ for all $R \in \hat{\mathcal{R}}$. Moreover, f is P-robust to misspecification with g' on $\mathcal{R}$. To see this, first note that if $f|_{\hat{\mathcal{R}}} \neq g|_{\hat{\mathcal{R}}}$, and $g|_{\hat{\mathcal{R}}} = g'|_{\hat{\mathcal{R}}}$, then $f \neq g'$. Moreover, if $\text{Im}(g|_{\hat{\mathcal{R}}}) \subseteq \text{Im}(f|_{\hat{\mathcal{R}}})$, and g' is equal to g on $\hat{\mathcal{R}}$ and equal to f outside $\hat{\mathcal{R}}$, then $\text{Im}(g') \subseteq \text{Im(f)}$. Moreover, we have assumed that $\text{Am}(f) \preceq P$ on $\mathcal{R}$. Finally, suppose that $f(R_1) = g'(R_2)$. If $R_2 \notin \hat{\mathcal{R}}$, then $g'(R_2) = f(R_2)$. Since $\text{Am}(f) \preceq P$ on $\mathcal{R}$, this implies that $R_1 \equiv_P R_2$. Next, if $R_1, R_2 \in \hat{\mathcal{R}}$, then $R_1 \equiv_P R_2$, since f is P-robust to misspecification with g on $\hat{\mathcal{R}}$. Finally, if $R_2 \in \hat{\mathcal{R}}$, $R_1 \notin \hat{\mathcal{R}}$, let R_3 be a reward function such that $R_3 \in \hat{\mathcal{R}}$ and $f(R_3) = g'(R_2)$. Since $\text{Im}(g|_{\hat{\mathcal{R}}}) \subseteq \text{Im}(f|_{\hat{\mathcal{R}}})$, such a reward function R_3 does exist. Since f is P-robust to misspecification with g on $\hat{\mathcal{R}}$, we have that $R_2 \equiv_P R_3$. Moreover, since $\text{Am}(f) \preceq P$ on $\mathcal{R}$, and $f(R_1) = f(R_3)$, we have that $R_1 \equiv_P R_3$. By transitivity, this implies that $R_1 \equiv_P R_2$. This covers all cases, which means that if $f(R_1) = g'(R_2)$ then $R_1 \equiv_P R_2$. Thus f is P-robust to misspecification with g' on $\mathcal{R}$. □

Theorem 119. *For any $\hat{\mathcal{R}} \subseteq \mathcal{R}$ and any pseudometric $d^{\mathcal{R}}$ on $\mathcal{R}$, if f is ϵ-robust to misspecification with g on $\mathcal{R}$ using $d^{\mathcal{R}}$, then f is ϵ-robust to misspecification with g on $\hat{\mathcal{R}}$ using $d^{\mathcal{R}}$, unless $f|_{\hat{\mathcal{R}}} = g|_{\hat{\mathcal{R}}}$. Similarly, if f is ϵ-robust to misspecification with g on $\hat{\mathcal{R}}$ using $d^{\mathcal{R}}$ then f is 2ϵ-robust to misspecification with g' on $\mathcal{R}$ using $d^{\mathcal{R}}$ for some g' where $g'|_{\hat{\mathcal{R}}} = g|_{\hat{\mathcal{R}}}$, unless there are $R_1, R_2 \in \mathcal{R}$ such that $f(R_1) = f(R_2)$ but $d^{\mathcal{R}}(R_1, R_2) > \epsilon$.*

Proof. For the first claim, suppose that f is ϵ-robust to misspecification with g on $\mathcal{R}$, and that $f|_{\hat{\mathcal{R}}} \neq g|_{\hat{\mathcal{R}}}$. Since f is ϵ-robust to misspecification with g on $\mathcal{R}$, we have that for all $R_1, R_2 \in \mathcal{R}$, if $f(R_1) = g(R_2)$ then $d^{\mathcal{R}}(R_1, R_2) \leq \epsilon$. Since $\hat{\mathcal{R}} \subseteq \mathcal{R}$, this directly implies that for all $R_1, R_2 \in \hat{\mathcal{R}}$, if $f(R_1) = g(R_2)$ then $d^{\mathcal{R}}(R_1, R_2) \leq \epsilon$. We also have that $\text{Im}(f) \subseteq \text{Im(g)}$, which directly implies that $\text{Im}(g|_{\hat{\mathcal{R}}}) \subseteq \text{Im}(f|_{\hat{\mathcal{R}}})$. Moreover, we have that for all $R_1, R_2 \in \mathcal{R}$, if $f(R_1) = g(R_2)$ then $d^{\mathcal{R}}(R_1, R_2) \leq \epsilon$. Since $\hat{\mathcal{R}} \subseteq \mathcal{R}$, this directly implies that for all $R_1, R_2 \in \hat{\mathcal{R}}$, if $f(R_1) = g(R_2)$ then $d^{\mathcal{R}}(R_1, R_2) \leq \epsilon$. We assume that $f|_{\hat{\mathcal{R}}} \neq g|_{\hat{\mathcal{R}}}$. Thus, f is ϵ-robust to misspecification with g on $\hat{\mathcal{R}}$.

For the second claim, suppose f is ϵ-robust to misspecification with g on $\hat{\mathcal{R}}$, and that for all $R_1, R_2 \in \mathcal{R}$, if $f(R_1) = f(R_2)$ then $d^{\mathcal{R}}(R_1, R_2) \leq \epsilon$. We construct a g' as follows; let $g'(R) = g(R)$ for all $R \in \hat{\mathcal{R}}$, and let $g'(R) = f(R)$ for all $R \notin \hat{\mathcal{R}}$. By construction, we have that $g(R) = g'(R)$ for all $R \in \hat{\mathcal{R}}$. Moreover, f is 2ϵ-robust to misspecification with g' on $\mathcal{R}$. To see this, first note that if $f|_{\hat{\mathcal{R}}} \neq g|_{\hat{\mathcal{R}}}$, and $g|_{\hat{\mathcal{R}}} = g'|_{\hat{\mathcal{R}}}$, then $f \neq g'$. Moreover, if $\text{Im}(g|_{\hat{\mathcal{R}}}) \subseteq \text{Im}(f|_{\hat{\mathcal{R}}})$, and g' is equal to g on $\hat{\mathcal{R}}$ and equal to f outside $\hat{\mathcal{R}}$, then $\text{Im}(g') \subseteq \text{Im(f)}$. Moreover, we have assumed that for all $R_1, R_2 \in \mathcal{R}$, if $f(R_1) = f(R_2)$ then $d^{\mathcal{R}}(R_1, R_2) \leq \epsilon \leq 2\epsilon$. Finally, suppose that $f(R_1) = g'(R_2)$. If $R_2 \notin \hat{\mathcal{R}}$, then $g'(R_2) = f(R_2)$. Since the upper diameter of $\text{Am}(f)$ on $\mathcal{R}$ is assumed to be at most ϵ, this implies that $d^{\mathcal{R}}(R_1, R_2) \leq \epsilon \leq 2\epsilon$. Next, if $R_1, R_2 \in \hat{\mathcal{R}}$, then $d^{\mathcal{R}}(R_1, R_2) \leq \epsilon \leq 2\epsilon$, since f is ϵ-robust to misspecification with g on $\hat{\mathcal{R}}$. Finally, if $R_2 \in \hat{\mathcal{R}}$, $R_1 \notin \hat{\mathcal{R}}$, let R_3 be a reward function such that $R_3 \in \hat{\mathcal{R}}$ and $f(R_3) = g'(R_2)$. Since $\text{Im}(g|_{\hat{\mathcal{R}}}) \subseteq \text{Im}(f|_{\hat{\mathcal{R}}})$, such a reward function R_3 does exist. Since f is ϵ-robust to misspecification with g on $\hat{\mathcal{R}}$, we have that $d^{\mathcal{R}}(R_2, R_3) \leq \epsilon$. Moreover, since the upper diameter of $\text{Am}(f)$ on $\mathcal{R}$ is at most ϵ, and $f(R_1) = f(R_3)$, we have that $d^{\mathcal{R}}(R_1, R_3) \leq \epsilon$. By the triangle inequality, this implies that $d^{\mathcal{R}}(R_1, R_2) \leq 2\epsilon$. This covers all cases, which means that if $f(R_1) = g'(R_2)$ then $d^{\mathcal{R}}(R_1, R_2) \leq 2\epsilon$. Thus f is 2ϵ-robust to misspecification with g' on $\mathcal{R}$ relative to the pseudometric $d^{\mathcal{R}}$. □

Theorem 118 says that if f is P-robust to misspecification with g on $\mathcal{R}$, then f is P-robust to misspecification with g on $\hat{\mathcal{R}}$ (unless $f = g$ on $\hat{\mathcal{R}}$). Thus, all *positive*

results generalise to arbitrary subsets $\hat{\mathcal{R}}$ of $\mathcal{R}$. The converse case is similar, but slightly more complicated; if f is P-robust to misspecification with g if and only if $g \in G$ for some G, then f is P-robust to misspecification with g' on $\hat{\mathcal{R}}$ if and only if g' behaves like some $g \in G$ for all $R \in \hat{\mathcal{R}}$. Roughly speaking, this means that if f is not robust to a given form of misspecification relative to $\mathcal{R}$, then this is still the case if we impose restrictions on the space of all reward functions, unless this restriction ensures that this misspecification has no impact on the generated data, or makes this misspecification equivalent to some other form of misspecification that f is robust to. Theorem 119 is analogous to Theorem 118.

To make this easier to understand, let us provide an example. Recall that the Boltzmann-rational behavioural model is $\mathrm{ORD}_{\tau,\gamma}$-robust to misspecification of the temperature parameter β, and no other misspecification. More precisely, relative to the set of all rewards $\mathcal{R}$, we have that $b_{\tau,\gamma,\beta}$, is $\mathrm{ORD}_{\tau,\gamma}$-robust to misspecification with g if and only if there is a function $\psi : \mathcal{R} \to \mathbb{R}$ such that $g(R) = b_{\tau,\gamma,\psi(R)}(R)$ and $b_{\tau,\gamma,\beta} \neq g$ (Theorem 89). In particular, $b_{\tau,\gamma,\beta}$ is not $\mathrm{ORD}_{\tau,\gamma}$-robust to misspecification of the discount parameter, γ. However, now suppose $|\mathcal{S}| \geq 3$, and that we pick a peculiar transition function τ, initial state distribution μ_0, and two states $s_0, s^\top \in \mathcal{S}$, $s_0 \neq s^\top$, such that

1. $\mathbb{P}_{S_0 \sim \mu_0}(S_0 = s_0) = 1$,
2. $\mathbb{P}_{S' \sim \tau(s_0,a)}(S' \in \{s_0, s^\top\}) = 0$ for all $a \in \mathcal{A}$, and
3. if $s \neq s_0$ then $\mathbb{P}_{S' \sim \tau(s,a)}(S' = s^\top) = 1$ for all $a \in \mathcal{A}$.

Moreover, let $\hat{\mathcal{R}}$ be the set of all reward functions R such that $R(s, a, s') = 0$ unless $s' = s^\top$ and $s \neq s^\top$, and such that $R(s, a_1, s^\top) = R(s, a_2, s^\top)$ for all $s \in \mathcal{S}$ and all $a_1, a_2 \in \mathcal{A}$. In other words, τ and μ_0 specify an environment in which $s^\top$ is a *terminal state* that any policy will enter after exactly 2 steps, and then stay in forever. Moreover, $\hat{\mathcal{R}}$ contains reward functions that only give a positive reward when the terminal state $s^\top$ is entered (from a state other than $s^\top$), and that give the same reward to any two transitions that enter $s^\top$ from the same state. This

means that any policy only can get non-zero reward on the second step of any run. Moreover, the value of a policy only depends on what action it takes in the initial state s_0, since all actions will have the same Q-values in all other states.

We now have that $b_{\tau,\gamma_1,\beta}$ is $\text{ORD}_{\tau,\gamma}$-robust to misspecification with $b_{\tau,\gamma_2,\beta}$ relative to $\hat{\mathcal{R}}$ for any γ_2 such that $\gamma_1 \neq \gamma_2$. To see this, note that changing γ in this environment has the same impact on the Boltzmann-rational policy as positive linear scaling of R, for all $R \in \hat{\mathcal{R}}$ (and that $\hat{\mathcal{R}}$ is closed under positive linear scaling). This means that, while $b_{\tau,\gamma,\beta}$ is not $\text{ORD}_{\tau,\gamma}$-robust to misspecification of γ on the space of all reward functions $\mathcal{R}$, it *is* robust to misspecification of γ relative to the set $\hat{\mathcal{R}}$ for this specific transition function and initial state distribution. However, the reason for this is that there for any γ_2 exists a β_2 such that $b_{\tau,\gamma_2,\beta}(R) = b_{\tau,\gamma_1,\beta_2}(R)$ for all $R \in \hat{\mathcal{R}}$. In other words, for this τ, μ_0, and $\hat{\mathcal{R}}$, misspecification of γ is equivalent to misspecification of β. Thus, restricting $\mathcal{R}$ to a smaller set $\hat{\mathcal{R}}$ can *in a sense* make a behavioural model f robust to forms of misspecification that it is not normally robust to. However, this can only happen if $\mathcal{R}$ is restricted in such a way that this misspecification becomes equivalent to a form of misspecification that f normally *is* robust to. This means that the setting where $\mathcal{R}$ is restricted ultimately is mostly similar to the setting where $\mathcal{R}$ is unrestricted, and that our results from Chapters 6 and 7 still are directly applicable.

However, this does not mean that it would be uninteresting to identify restrictions on $\mathcal{R}$ that make certain kinds of misspecification less problematic. For example, Theorem 106 shows that a wide range of behavioural models in principle can be highly sensitive to arbitrarily small misspecification of the discount parameter, γ. However, although it is rather contrived, our example above shows that even arbitrary misspecification of γ can be entirely unproblematic in certain environments, for certain restricted sets of reward functions. This means that the negative effects of misspecification, at least in some cases, can be mitigated by incorporating prior knowledge about the true reward function. By studying the effects of different types of restrictions on $\mathcal{R}$, we might get a better understanding of what forms of prior knowledge can help with what forms of misspecification.

For example, Theorem 105 shows that if f_τ is invariant to S'-redistribution, and $\tau_1 \neq \tau_2$, then f_{τ_1} is not ϵ-robust to misspecification with f_{τ_2} as measured by $d^{\text{STARC}}_{\tau_1,\gamma}$, for any $\epsilon < 0.5$. The reason for this, intuitively, is that there are reward functions that differ by S'-redistribution under τ_2, but which have a large distance under $d^{\text{STARC}}_{\tau_1,\gamma}$. However, suppose we restrict $\mathcal{R}$ to the set $\hat{\mathcal{R}}$ of reward functions such that if $R \in \hat{\mathcal{R}}$, then for all s, a, s_1, s_2, we have that $R(s, a, s_1) = R(s, a, s_2)$. In that case there are no reward functions in $\hat{\mathcal{R}}$ that differ by S'-redistribution, which suggests that misspecification of τ might be less problematic relative to this set of rewards. Similarly, Theorem 106 says that if f_γ is invariant to potential shaping, and $\gamma_1 \neq \gamma_2$ (and τ is non-trivial), then f_{γ_1} is not ϵ-robust to misspecification with f_{γ_2} as measured by $d^{\text{STARC}}_{\tau,\gamma_1}$, for any $\epsilon < 0.5$. Again, the reason for this is, intuitively, that there are reward functions that differ by potential shaping under γ_2, but which have a large distance under $d^{\text{STARC}}_{\tau,\gamma_1}$. However, suppose we restrict $\mathcal{R}$ to the set $\hat{\mathcal{R}}$ of reward functions that only reward a single transition (or the set of reward functions that only are non-zero in a single state, etc). In that case there are no reward functions in $\hat{\mathcal{R}}$ that differ by potential shaping, which suggests that misspecification of γ might be less problematic relative to this set of rewards. Investigating the relationship between misspecification robustness and restrictions on $\mathcal{R}$ may thus be an interesting and fruitful direction for future work.

We think it is also worth remarking on Theorem 104, which says that no continuous behavioural model is ϵ/δ-separating relative to $d^{\text{STARC}}_{\tau,\gamma}$. This also means that no continuous behavioural model is perturbation robust relative to $d^{\text{STARC}}_{\tau,\gamma}$. The proof of Theorem 104 finds a specific counterexample in the vicinity of the zero reward, R_0, and these kinds of counterexamples could be ruled out by imposing certain restrictions on $\hat{\mathcal{R}}$. A natural suggestion might be to require that each reward function in $\hat{\mathcal{R}}$ should be normalised, or perhaps that each reward function in $\hat{\mathcal{R}}$ should have a certain minimum L_2-norm, etc. Our next result shows that such restrictions are insufficient to mitigate Theorem 104:

Theorem 120. *Let $d^{\mathcal{R}}$ be $d^{\text{STARC}}_{\tau,\gamma}$, and let d^{Π} be a pseudometric on Π which satisfies the condition that for all δ_1 there exists a δ_2 such that if $L_2(\pi_1, \pi_2) < \delta_2$*

then $d^{\Pi}(\pi_1, \pi_2) < \delta_1$. *Let* c *be any positive constant, and let* $\hat{\mathcal{R}}$ *be a set of reward functions such that if* $L_2(R) = c$ *then* $R \in \hat{\mathcal{R}}$. *Let* $f : \hat{\mathcal{R}} \to \Pi$ *be continuous. Then* f *is not* ϵ/δ*-separating for any* $\epsilon < 1$ *or* $\delta > 0$ *on* $\hat{\mathcal{R}}$.

Proof. Let R be a non-trivial reward function that is orthogonal to all trivial reward functions. Since the set of all trivial reward functions form a linear subspace (Proposition 33 and Theorem 40), such a reward function R exists. Note that R must not necessarily be contained in $\hat{\mathcal{R}}$.

Since R is non-trivial, we have that $d^{\text{STARC}}_{\tau,\gamma}(\epsilon R, -\epsilon R) = 1$ for any positive constant ϵ. Next, let R_Φ be some potential-shaping reward function such that $L_2(\epsilon R + R_\Phi) = c$. Since potential shaping does not change the ordering of policies, we have that $d^{\text{STARC}}_{\tau,\gamma}(\epsilon R + R_\Phi, -\epsilon R + R_\Phi) = 1$. Moreover, since R_Φ is trivial, we have that both ϵR and $-\epsilon R$ are orthogonal to R_Φ, and so $L_2(-\epsilon R + R_\Phi) = c$ as well. Since $L_2(\epsilon R + R_\Phi) = L_2(\epsilon R + R_\Phi) = c$, we have that both $\epsilon R + R_\Phi$ and $-\epsilon R + R_\Phi$ are contained in $\hat{\mathcal{R}}$.

Let δ_1 be any positive constant. By assumption, there exists a δ_2 such that if $L_2(\pi_1 - \pi_2) < \delta_2$ then $d^{\Pi}(\pi_1, \pi_2) < \delta_1$. Moreover, since f is continuous, there exists an ϵ_1 such that if $L_2(R_1, R_2) < \epsilon_1$, then $L_2(f(R_1), f(R_2)) < \delta_2$. Next, note that by making ϵ sufficiently small, we can ensure that the L_2-distance between $\epsilon R + R_\Phi$ and $\epsilon R + R_\Phi$ is arbitrarily small (and, in particular, less than ϵ_1).

Thus, for any positive δ there exist reward functions $\epsilon R + R_\Phi$ and $-\epsilon R + R_\Phi$ that are both contained in $\hat{\mathcal{R}}$, such that $d^{\Pi}(f(\epsilon R + R_\Phi), f(-\epsilon R + R_\Phi)) < \delta$, and such that $d^{\text{STARC}}_{\tau,\gamma}(\epsilon R + R_\Phi, -\epsilon R + R_\Phi) = 1$. Thus f is not ϵ/δ-separating for any $\delta > 0$ and any $\epsilon < 1$. □

The reason for why Theorem 120 is true is that there are reward functions which are "nearly" trivial, in the sense that $\max_\pi J(\pi) \approx \min_\pi J(\pi)$, but which nonetheless have a large norm — we can find such reward functions by applying potential shaping and S'-redistribution to rewards in the vicinity of R_0. This does not necessarily mean that we cannot circumvent Theorem 104 by finding an appropriate restriction on $\mathcal{R}$, but it does mean that simple normalisation is insufficient for doing so.

8.4 Making the Analysis Probabilistic

The definitions we have given in Chapter 3 provide what is essentially a worst-case analysis, in the sense that they require each condition to hold for *all* reward functions. However, in certain cases, we may know that $R^\star$ is sampled from a particular distribution $\mathcal{D}$ over $\mathcal{R}$. In those cases, it may be more relevant to know whether the learnt reward function R_H is similar (in a relevant sense) to the true reward $R^\star$ with *high probability*. In this section, we will discuss this generalisation.

To make this setting a bit more formal, we may assume that we have two reward objects $f, g : \mathcal{R} \to X$ and a distribution $\mathcal{D}$ over $\mathcal{R}$, that $R^\star$ is sampled from $\mathcal{D}$, and that the learning algorithm $\mathcal{L}$ observes the data given by $g(R^\star)$. We then assume that $\mathcal{L}$ returns a reward function R_H such that $f(R_H) = g(R^\star)$, and that $\mathcal{L}$ selects among all such reward functions using some (potentially nondeterministic) inductive bias. We can then ask whether or not $R_H \equiv_P R^\star$ with probability at least $1-\delta$, or $d^{\mathcal{R}}(R^\star, R_h) \leq \epsilon$ with probability at least $1-\delta$, for some δ and ϵ, and some partition P or pseudometric $d^{\mathcal{R}}$ on $\mathcal{R}$. As usual, if $f \neq g$, then f is misspecified, and otherwise f is correctly specified. Moreover, if the learnt reward R_H will be used to compute some object $h : \mathcal{R} \to Y$, then we can set $P = \mathrm{Am}(h)$.

To some extent, our analysis in Section 8.3 can be used to understand this setting as well. In particular, suppose we pick a set $\hat{\mathcal{R}}$ of "likely" reward functions such that $\mathbb{P}_{R\sim\mathcal{D}}(R \in \hat{\mathcal{R}}) \geq 1-\delta$, and such that the learning algorithm $\mathcal{L}$ will return a reward function $R_h \in \hat{\mathcal{R}}$ if there exists a reward function $R_h \in \hat{\mathcal{R}}$ such that $f(R_h) = g(R^\star)$. Then if f is P-robust to misspecification with g on $\hat{\mathcal{R}}$, we have that $\mathcal{L}$ will learn a reward function R_H such that $R_H \equiv_P R^\star$ with probability at least $1-\delta$. Similarly, if f is ϵ-robust to misspecification with g on $\hat{\mathcal{R}}$, then $\mathcal{L}$ will learn a reward function R_H such that $d^{\mathcal{R}}(R^\star, R_H) \leq \epsilon$ with probability at least $1-\delta$.

So, for example, suppose $\hat{\mathcal{R}}$ is the set of all reward functions that have "low complexity", for some complexity measure and complexity threshold. The above argument then informally tells us that if the true reward function is likely to have low complexity, and if $\mathcal{L}$ will attempt to fit a low-complexity reward function to its training data, then the learnt reward function will be close to the true reward

function with high probability, as long as f is robust to misspecification with g on the set of all low-complexity reward functions. Similarly, if the true reward function is likely to be sparse, and $\mathcal{L}$ will attempt to fit a sparse reward function to its training data, then we may let $\hat{\mathcal{R}}$ be equal to the set of all sufficiently sparse reward functions, and so on.

Thus, while our definitions in Chapter 3 give us a worst-case formalisation of ambiguity and misspecification robustness, it is relatively straightforward to carry out a more probabilistic analysis within the same framework. In particular, our results in Section 8.3 directly provide us with results about the probabilistic setting as well. However, this analysis could probably be extended with more specific results. Doing so is out of scope for this dissertation, but it may be an interesting direction for future work.

8.5 Stronger Equivalence Conditions

We have introduced three primary methods for characterising the difference between two reward functions; the equivalence relations given by $\mathrm{ORD}_{\tau,\gamma}$ and $\mathrm{OPT}_{\tau,\gamma}$, and the pseudometrics given by STARC. Intuitively, $\mathrm{ORD}_{\tau,\gamma}$ considers R_1 and R_2 to be equivalent if they have the same ordering of policies under τ and γ, and $\mathrm{OPT}_{\tau,\gamma}$ considers R_1 and R_2 to be equivalent if they have the same optimal policies under τ and γ. Similarly, STARC metrics consider R_1 and R_2 to be similar if they have a similar ordering of policies under τ and γ.

It is worth noting that all of these are parameterised in terms of τ and γ; this means that $\mathrm{ORD}_{\tau,\gamma}$ only requires R_1 and R_2 to have the same ordering of policies in one particular environment, and so on. Moreover, two reward functions can have the same policy order in one environment, without having the same policy order in a different environment. To see this, note that Theorem 40 says that R_1 and R_2 have the same ordering of policies if and only if R_1 and R_2 differ by potential shaping, S'-redistribution, and positive linear scaling, and these transformations are parameterised by τ and γ. Thus, while $R_1 \equiv_{\mathrm{ORD}_{\tau,\gamma}} R_2$ ensures

that R_1 and R_2 are equivalent under τ and γ, it does not guarantee this for other transition functions or discounts.

To guarantee robust transfer learning, one may therefore wish to consider stronger equivalence relations. Unfortunately, the reward learning methods that we consider are unable to ensure that the learnt reward function has such guarantees (see Section 5.4), which makes it redundant to consider stronger equivalence relations or metrics than those we have already presented. Nonetheless, for the sake of completeness, we will briefly discuss how to create stronger equivalence relations on $\mathcal{R}$. First, let $\text{ORD}_\star$ be the equivalence relation according to which $R_1 \equiv_{\text{ORD}_\star} R_2$ if and only if, for any state s, we have that

$$\mathbb{E}_{\xi\sim D_1}\left[G_1(\xi)\right] \geq \mathbb{E}_{\xi\sim D_2}\left[G_1(\xi)\right] \iff \mathbb{E}_{\xi\sim D_1}\left[G_2(\xi)\right] \geq \mathbb{E}_{\xi\sim D_2}\left[G_2(\xi)\right]$$

for all distributions D_1, D_2 over trajectories that start in s. This means that if $R_1 \equiv_{\text{ORD}_\star} R_2$, then R_1 and R_2 induce the same preferences over all policies for all transition functions τ — indeed, this will hold even if we permit τ to be non-Markovian, etc. To characterise $\text{ORD}_\star$, we will make use of the following lemma:

Lemma 121. *Let $\Xi = (\mathcal{S}\times\mathcal{A})^\omega$ be the set of all trajectories, and let $d : \Xi \times \Xi \to \mathbb{R}$ be the function given by $d(\xi_1, \xi_2) = 0$ if $\xi_1 = \xi_2$, and else $d(\xi_1, \xi_2) = \frac{1}{e^t}$, where t is the smallest index on which ξ_1 and ξ_2 differ. Then (Ξ, d) is a compact metric space.*

Proof. We must first show that d is a metric, which requires showing that it satisfies the following:

1. Indiscernibility of Identicals: $d(\xi_1, \xi_2) = 0$ if $\xi_1 = \xi_2$.
2. Identity of Indiscernibles: $d(\xi_1, \xi_2) = 0$ only if $\xi_1 = \xi_2$.
3. Positivity: $d(\xi_1, \xi_2) \geq 0$.
4. Symmetry: $d(\xi_1, \xi_2) = d(\xi_2, \xi_1)$.
5. Triangle Inequality: $d(\xi_1, \xi_3) \leq d(\xi_1, \xi_2) + d(\xi_2, \xi_3)$.

It is straightforward to see that 1-4 hold. For 5, let t be the smallest index on which ξ_1 and ξ_3 differ. Note that if $d(\xi_1, \xi_3) > d(\xi_1, \xi_2)$ and $d(\xi_1, \xi_3) > d(\xi_2, \xi_3)$, then it must be the case that $\xi_1[i] = \xi_2[i]$ for all $i \leq t$, and that $\xi_2[i] = \xi_3[i]$ for all $i \leq t$. However, this is a contradiction, since it would imply that $\xi_1[t] = \xi_3[t]$. Thus either $d(\xi_1, \xi_3) \leq d(\xi_1, \xi_2)$ or $d(\xi_1, \xi_3) \leq d(\xi_2, \xi_3)$, which in turn implies that $d(\xi_1, \xi_3) \leq d(\xi_1, \xi_2) + d(\xi_2, \xi_3)$.

Thus d is a metric, which means that (Ξ, d) is a metric space. Next, we will prove that (Ξ, d) is compact. We will do this by showing that (Ξ, d) is totally bounded and complete.

To see that (Ξ, d) is totally bounded, let ϵ be an arbitrary positive real number, and let $t = \ln(1/\epsilon)$, so that $\epsilon = 1/e^t$. Moreover, let s and a be an arbitrary state and action, and let $\hat{\Xi}$ be the set of all trajectories ξ such that $\xi[i] = \langle s, a \rangle$ for all $i > t$ (but which may include arbitrary transitions before time t). Now $\hat{\Xi}$ is finite, and for every trajectory ξ_1 there is a trajectory $\xi_2 \in \hat{\Xi}$ such that $d(\xi_1, \xi_2) \leq \epsilon$ (given by letting $\xi_2[i] = \xi_1[i]$ for all $i \leq t$). Thus, for every $\epsilon > 0$, we have that (Ξ, d) has a finite cover. This means that (Ξ, d) is totally bounded.

To see that (Ξ, d) is complete, let $\{\xi_i\}_{i=0}^{\infty}$ be a Cauchy sequence. This implies that for every $\epsilon > 0$ there is a positive integer N such that for all $n, m \geq N$ we have $d(\xi_n, \xi_m) < \epsilon$. In our case, this means that there, for each time t is a positive integer N such that for all $n, m \geq N$, we have that $\xi_n[i] = \xi_m[i]$ for all $i \leq t$. We can thus define a trajectory ξ_∞ by letting $\xi_\infty[i] = \langle s, a \rangle$ if there is an N such that, for all $n \geq N$, we have that $\xi_n[i] = \langle s, a \rangle$. Now $\lim_{i \to \infty} \{\xi_i\}_{i=0}^{\infty} = \xi_\infty$, and $\xi_\infty \in (\Xi, d)$. Thus every Cauchy sequence in (Ξ, d) has a limit that is also in (Ξ, d), and so (Ξ, d) is complete.

Every metric space which is totally bounded and complete is compact. Thus, (Ξ, d) is a compact metric space. □

We can now characterise $\text{ORD}_\star$ as follows:

Proposition 122. *$R_1 \equiv_{\text{ORD}_\star} R_2$ if and only if R_1 and R_2 differ by potential shaping and positive linear scaling.*

Proof. Fix a state s. It is now straightforward that

$$\mathbb{E}_{\xi\sim D_1}[G_1(\xi)] \geq \mathbb{E}_{\xi\sim D_2}[G_1(\xi)] \iff \mathbb{E}_{\xi\sim D_1}[G_2(\xi)] \geq \mathbb{E}_{\xi\sim D_2}[G_2(\xi)]$$

for all distributions D_1, D_2 over trajectories that start in s, if and only if G_1 and G_2 differ by an affine transformation for all trajectories starting in s. Moreover, this corresponds exactly to potential shaping and positive linear scaling of R, as per Proposition 31 and 32. Furthermore, since this condition holds for all $s \in \mathcal{S}$ whenever it holds for one $s \in \mathcal{S}$, this means that $R_1 \equiv_{\text{ORD}_\star} R_2$ if and only if R_1 and R_2 differ by potential shaping and positive linear scaling.

We will next show rigorously that

$$\mathbb{E}_{\xi\sim D_1}[G_1(\xi)] \geq \mathbb{E}_{\xi\sim D_2}[G_1(\xi)] \iff \mathbb{E}_{\xi\sim D_1}[G_2(\xi)] \geq \mathbb{E}_{\xi\sim D_2}[G_2(\xi)]$$

for all distributions D_1, D_2 over trajectories that start in s, if and only if G_1 and G_2 differ by an affine transformation for all trajectories starting in s. For the first direction, suppose there is an $a \in \mathbb{R}^+$ and a $b \in \mathbb{R}$ such that $G_2(\xi) = a \cdot G_1(\xi) + b$ for all trajectories ξ that start in s. Then $\mathbb{E}_{\xi\sim D}[G_2(\xi)] = a \cdot \mathbb{E}_{\xi\sim D}[G_1(\xi)] + b$ for all distributions D over trajectories that start in s, by the linearity of expectation. This in turn implies that

$$\mathbb{E}_{\xi\sim D_1}[G_1(\xi)] \geq \mathbb{E}_{\xi\sim D_2}[G_1(\xi)] \iff \mathbb{E}_{\xi\sim D_1}[G_2(\xi)] \geq \mathbb{E}_{\xi\sim D_2}[G_2(\xi)]$$

for all distributions D_1, D_2 over trajectories that start in s, since

$$\mathbb{E}_{\xi\sim D_1}[G_1(\xi)] \geq \mathbb{E}_{\xi\sim D_2}[G_1(\xi)] \iff a \cdot \mathbb{E}_{\xi\sim D_1}[G_1(\xi)] + b \geq a \cdot \mathbb{E}_{\xi\sim D_2}[G_1(\xi)] + b.$$

For the other direction, suppose

$$\mathbb{E}_{\xi\sim D_1}[G_1(\xi)] \geq \mathbb{E}_{\xi\sim D_2}[G_1(\xi)] \iff \mathbb{E}_{\xi\sim D_1}[G_2(\xi)] \geq \mathbb{E}_{\xi\sim D_2}[G_2(\xi)]$$

for all distributions D_1, D_2 over trajectories that start in s. Next, let $\Xi = (\mathcal{S}\times\mathcal{A})^\omega$ be the set of all trajectories, and let $d : \Xi \times \Xi \to \mathbb{R}$ be the function given by $d(\xi_1, \xi_2) = \frac{1}{e^t}$, where t is the smallest index on which ξ_1 and ξ_2 differ, or 0 if $\xi_1 = \xi_2$. As per Lemma 121, (d, Ξ) is a compact metric space. Moreover, it is easy to see that

this still holds if we restrict Ξ to the set of trajectories Ξ_s that start in s, and that G_1 is continuous with respect to the metric d. As per the extreme value theorem, this implies that there are two trajectories $\xi_1, \xi_2 \in \Xi_s$ such that $G_1(\xi_1) \leq G_1(\xi) \leq G_1(\xi_2)$ for all $\xi \in \Xi_s$. Next, if $G_1(\xi_1) = G_1(\xi_2)$, then $G_1(\xi) = G_1(\xi')$ for all ξ, ξ', in which case $G_2(\xi) = G_2(\xi')$ for all ξ, ξ' as well. In this case, simply let $a = 1$ and $b = G_2(\xi) - G_1(\xi)$. Otherwise, let

$$a = \frac{G_2(\xi_2) - G_2(\xi_1)}{G_1(\xi_2) - G_1(\xi_1)},$$

and let $b = G_2(\xi_1) - a \cdot G_1(\xi_1)$. By rearranging, we see that $G_2(\xi_1) = a \cdot G_1(\xi_1) + b$ and $G_2(\xi_2) = a \cdot G_1(\xi_2) + b$. Moreover, let ξ be an arbitrary trajectory in Ξ_s. The intermediate value theorem now implies that there is a $p \in [0, 1]$ such that

$$(1 - p) \cdot G_1(\xi_1) + p \cdot G_1(\xi_2) = G_1(\xi).$$

Moreover, since $(1 - p) \cdot G_1(\xi_1) + p \cdot G_1(\xi_2)$ grows monotonically in p, this value must be unique. We can now consider the distribution D_1 over Ξ_s that returns ξ_2 with probability p, and ξ_1 otherwise, and the distribution D_2 that returns ξ with probability 1. This means that $\mathbb{E}_{\xi \sim D_1}[G_1(\xi)] = \mathbb{E}_{\xi \sim D_2}[G_1(\xi)]$, and so $\mathbb{E}_{\xi \sim D_1}[G_2(\xi)] = \mathbb{E}_{\xi \sim D_2}[G_2(\xi)]$. In other words,

$$(1 - p) \cdot G_2(\xi_1) + p \cdot G_2(\xi_2) = G_2(\xi),$$

which in turn implies that

$$(1 - p) \cdot (a \cdot G_1(\xi_1) + b) + p \cdot (a \cdot G_1(\xi_2) + b) = G_2(\xi).$$

By rearranging, and simplifying, we get that

$$a \cdot ((1 - p) \cdot G_1(\xi_1) + p \cdot G_1(\xi_2)) + b = G_2(\xi)$$

and so $G_2(\xi) = a \cdot G_1(\xi) + b$. Since ξ was chosen arbitrarily, this holds for all $\xi \in \Xi_s$. This completes the other direction. □

Moreover, we would like to remark on a minor subtlety. One might expect that the equivalence relation $\equiv_\Omega$ according to which $R_1 \equiv_\Omega R_2$ if and only if $R_1 \equiv_{\mathrm{ORD}_{\tau,\gamma}} R_2$ for all τ, is the same as $\mathrm{ORD}_\star$. However, this is not the case. To see this, consider the rewards R_1, R_2 where $R_1(s_1, a_1, s_1) = 1$, $R_1(s_1, a_1, s_2) = 0.5$, $R_2(s_1, a_1, s_1) = 0.5$, and $R_2(s_1, a_1, s_2) = 1$, and where R_1 and R_2 are 0 for all other transitions. Now R_1 and R_2 do not differ by potential shaping and linear scaling, yet they have the same policy order for all τ. The reason for this is that $\mathrm{ORD}_\star$ considers all distributions over ξ, and that not all of these distributions can be realised by some policy π and some transition function τ. Characterising $\equiv_\Omega$ would therefore require a more careful analysis of the impact of changing the transition function τ.

Information is closely associated with uncertainty. The information I obtain when you say something to me corresponds to the amount of uncertainty I had, previous to your speaking, of what you were going to say.

— Claude Shannon, 1953.

9 Discussion

In this chapter, we discuss the impact and significance of our results, their limitations, and how they may be extended and expanded upon in future work.

9.1 Impact and Significance

We have shown that both the partial identifiability as well as the misspecification robustness of behavioural models in IRL can be both quantified and understood. Specifically, we have fully characterised the partial identifiability (or the *ambiguity*) of the reward function under the three standard behavioural models, and we have fully characterised all forms of misspecification that these behavioural models are robust to. Moreover, we have shown that these results can be used to gain an intuitive insight into the practical consequences of partial identifiability and misspecification in IRL.

Our results show that the ambiguity of the reward function under the Boltzmann-rational model and the MCE model is low as long as the learnt reward function is evaluated in the training environment, whereas the ambiguity under the optimality model is larger. Moreover, and perhaps surprisingly, we have shown that each of these models can be too ambiguous to guarantee that the learnt reward function

robustly leads to desirable behaviour in new environments (namely, environments where the transition function τ or discount γ differ from the training environment).

We have shown that the optimality model lacks robustness to any kind of misspecification, whereas both the Boltzmann-rational model and the MCE model are robust to several forms of misspecification, where the exact forms of misspecification they are robust to depends on how we quantify the error in the learnt reward function. However, we have shown that none of these behavioural models are robust to even slight misspecification of the transition function τ or the discount function γ. Moreover, we have shown that very minimal assumptions about the behavioural model are needed to obtain this result, which means that new behavioural models are likely to also have this property. We find this quite surprising, as in the reinforcement learning literature the discount γ is typically selected in a somewhat arbitrary way, and it can often be difficult to establish post-facto which γ was used to compute a given policy. The fact that τ must be specified correctly is somewhat less surprising (considering, for instance, the examples discussed by Freedman, Shah, and A. Dragan, 2020), yet important to have established. We have also shown that none of these behavioural models are robust to arbitrarily small perturbations of the observed policy. We have similarly needed very minimal assumptions about the behavioural model to obtain this result, which means that this result also is likely to generalise to new behavioural models.

In addition to these contributions, we have also derived several powerful formal tools that can be used in the analysis of reward learning algorithms. First of all, in Chapter 4, we have provided a wide range of useful results about the properties of reward functions. Notably, we have introduced STARC metrics, shown that these pseudometrics induce both an upper and a lower bound on worst-case regret (see Definitions 43 and 44), and that they are unique in doing so. Thus, STARC metrics are an appropriate tool for analysing the properties and performance of reward learning algorithms. We have also provided necessary and sufficient conditions that describe when two reward functions have the same optimal policies, or the same ordering of policies, and we have elucidated the properties of many important forms

of reward transformations. In addition to this, we have provided several powerful lemmas in Chapter 3, that are useful for proving results about partial identifiability and misspecification robustness. We expect these results and contributions to be useful for further theoretical analysis of reward learning algorithms, beyond the analysis that we have carried out in this paper.

Our analysis provides a first step towards answering the more general question of how sensitive IRL is to misspecification of the behavioural model. Our results show that a very wide range of behavioural models — including all the three behavioural models that are most common in the current IRL literature — can be highly sensitive to some types of misspecification, namely misspecification of the transition function τ or discount factor γ, or perturbations of the observed policy. These results indicate that IRL in general can be highly sensitive to misspecification of the behavioural model. This provides a cautionary lesson on the prospects of IRL as a tool for accurate preference elicitation. The relationship between human preferences and human behaviour is very complex, and while it is certainly possible to create increasingly accurate models of human behaviour, it will never be realistically possible to create a behavioural model that is completely free forms of misspecification. Therefore, if IRL is unable to guarantee accurate inferences under even mild misspecification of the behavioural model, then we should expect it to be very difficult to guarantee that IRL reliably will produce accurate inferences in real-world situations. Our results suggest that IRL should be used cautiously, and that the learnt reward functions should be carefully evaluated (as done by e.g. Michaud, Gleave, and Russell, 2020; Jenner and Gleave, 2022). This also means that we need IRL algorithms that are specifically designed to be more robust to misspecification, such as e.g. that proposed by Viano et al. (2021). It may also be fruitful to combine IRL with other data sources, as done by e.g. Ibarz et al., 2018b, or consider policy optimisation algorithms that conservatively assume that the reward may be misspecified, as done by e.g. Krakovna, Orseau, Kumar, et al. (2018), Krakovna, Orseau, Ngo, et al. (2020), Turner, Ratzlaff, and Tadepalli (2020), and Griffin et al. (2022).

9.2 Limitations and Further Work

There are several ways to extend our work. First of all, our analysis has primarily focused on the three behavioural models that are most common in the current IRL literature (namely optimality, Boltzmann-rationality, and MCE optimality). One way to extend our work is to consider broader classes of behavioural models, or behavioural models that are more realistic. For example, there is an extensive body of work in the behavioural sciences that suggests that human behaviour (and that of many other animals) are better modelled using hyperbolic discounting, rather than exponential discounting for example, see Thaler, 1981; Mazur, 1987; Green and Myerson, 1996; Kirby, 1997; Frederick, Loewenstein, and O'Donoghue, 2002. However, the three standard behavioural models are all based on exponential discounting. It would therefore be interesting to extend our analysis to behavioural models that are based on hyperbolic discounting (or other kinds of discounting). Similarly, humans typically exhibit risk-averse behaviour, according to which losses are given a greater weight than gains, and this is not modelled by any of the three standard behavioural models. Therefore, it would also be interesting to extend our analysis to behavioural models that incorporate current models of human risk-aversion, such as *prospect theory* (Kahneman and Tversky, 1979). Alternatively, our analysis could also be extended by deriving results that apply to very wide classes of behavioural models, obtained by raising minimal assumptions (as we do in e.g. Sections 5.4 and 6.2).

Another way to extend our work is to consider other equivalence relations or other pseudometrics on $\mathcal{R}$. Much of our analysis has been based on the equivalence relations given by $\mathrm{ORD}_{\tau,\gamma}$ and $\mathrm{OPT}_{\tau,\gamma}$, as well as on STARC metrics. Note that we have shown that any pseudometric on $\mathcal{R}$ that gives rise to both an upper and a lower bound on worst-case regret must be bilipschitz equivalent to STARC metrics, so these pseudometrics must have a degree of canonicity. However, our definition of regret is quite strong: it may thus be possible to create more permissive pseudometrics, by allowing them to induce guarantees that are weaker than a

worst-case regret bound. There may also be other interesting equivalence relations on $\mathcal{R}$ that it would be worthwhile to study, besides $\text{ORD}_{\tau,\gamma}$ and $\text{OPT}_{\tau,\gamma}$.

Next, our work has assumed that the environment is described by a single-agent MDP. An interesting extension would be to considering more general classes of environments, such as multi-agent environments, environments with partial observability, environments with non-Markovian dynamics, or environments where the actions of the agent may be predicted in advance (as done by e.g. Bell et al., 2021). We have also assumed that $\mathcal{S}$ and $\mathcal{A}$ are finite. However, this does not hold in continuous environments, which are common in practice, and so it may be interesting to lift this assumption as well. Note that such extensions also would require the results in Chapter 4 to be generalised.

Another interesting direction for future work is to extend our analysis in Chapter 8, by more carefully considering the consequences of imposing restrictions on the set of rewards $\mathcal{R}$, or the consequences of using a *probability distribution* over $\mathcal{R}$, and demanding that the learnt reward R_H is close to the true reward $R^\star$ *with high probability*. We provide a range of results regarding these settings in Chapter 8. However, these results are somewhat preliminary, and it seems likely that it would be possible to draw additional interesting conclusions if these settings are explored more extensively. In Chapter 8, we also provide a more detailed discussion of a few different ways in which this could be done.

Furthermore, our analysis primarily concerns the asymptotic behaviour of IRL algorithms, in the limit of infinite data. Thus an interesting extension could study the properties of IRL algorithms in the case of finite data (as done by e.g. Metelli, Lazzati, and Restelli, 2023). Finally, whilst our analysis has been theoretical, it could be insightful to study the impact of misspecification in IRL from an empirical angle (as done by e.g. Chan, Critch, and A. Dragan, 2021).

References

Abbeel, Pieter, Adam Coates, and Andrew Y Ng (2010). "Autonomous Helicopter Aerobatics Through Apprenticeship Learning". In: *The International Journal of Robotics Research* 29.13, pp. 1608–1639. DOI: 10.1177/0278364910371999.

Abel, David (2022). *A Theory of Abstraction in Reinforcement Learning*. arXiv: 2203.00397 [cs.LG]. URL: https://arxiv.org/abs/2203.00397.

Abel, David, D. Ellis Hershkowitz, and Michael L. Littman (2017). *Near Optimal Behavior via Approximate State Abstraction*. arXiv: 1701.04113 [cs.LG]. URL: https://arxiv.org/abs/1701.04113.

Abel, David, Nate Umbanhowar, et al. (Aug. 2020). "Value Preserving State-Action Abstractions". In: *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*. Ed. by Silvia Chiappa and Roberto Calandra. Vol. 108. Proceedings of Machine Learning Research. PMLR, pp. 1639–1650. URL: https://proceedings.mlr.press/v108/abel20a.html.

Adams, Stephen, Tyler Cody, and Peter A Beling (Aug. 2022). "A survey of inverse reinforcement learning". en. In: *Artif. Intell. Rev.* 55.6, pp. 4307–4346.

Allais, M. (1953). "Le Comportement de l'Homme Rationnel devant le Risque: Critique des Postulats et Axiomes de l'Ecole Americaine". In: *Econometrica* 21.4, pp. 503–546. ISSN: 00129682, 14680262. URL: http://www.jstor.org/stable/1907921 (visited on 12/15/2023).

Armstrong, Stuart and Sören Mindermann (2019). *Occam's razor is insufficient to infer the preferences of irrational agents*. arXiv: 1712.05812 [cs.AI].

Arora, Saurabh and Prashant Doshi (2020). *A Survey of Inverse Reinforcement Learning: Challenges, Methods and Progress*. arXiv: 1806.06877 [cs.LG].

Bell, James et al. (2021). "Reinforcement Learning in Newcomblike Environments". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato et al. Vol. 34. Curran Associates, Inc., pp. 22146–22157. URL: https://proceedings.neurips.cc/paper_files/paper/2021/file/b9ed18a301c9f3d183938c451fa183df-Paper.pdf.

Cai, Xin-Qiang et al. (2022). *Seeing Differently, Acting Similarly: Heterogeneously Observable Imitation Learning*. arXiv: 2106.09256 [cs.LG]. URL: https://arxiv.org/abs/2106.09256.

Cao, Haoyang, Samuel N. Cohen, and Lukasz Szpruch (2021). "Identifiability in Inverse Reinforcement Learning". In: *arXiv preprint* arXiv:2106.03498 [cs.LG].

Chan, Lawrence, Andrew Critch, and Anca Dragan (2021). "Human irrationality: both bad and good for reward inference". In: *arXiv preprint arXiv:2111.06956*. arXiv: 2111.06956 [cs.LG].

Chernozhukov, Victor, Han Hong, and Elie Tamer (2007). "Estimation and Confidence Regions for Parameter Sets in Econometric Models". In: *Econometrica* 75.5, pp. 1243–1284. ISSN: 00129682, 14680262. URL: http://www.jstor.org/stable/4502031 (visited on 04/14/2025).

Clark, Jack and Dario Amodei (2016). *Faulty Reward Functions in the Wild*. OpenAI Codex https://openai.com/blog/faulty-reward-functions/.

Dvijotham, Krishnamurthy and Emanuel Todorov (June 2010). "Inverse Optimal Control with Linearly-Solvable MDPs". In: *Proceedings of the 27th International Conference on Machine Learning*. Haifa, Israel: Omnipress, Madison, Wisconsin, USA, pp. 335–342.

Ellsberg, Daniel (Nov. 1961). "Risk, Ambiguity, and the Savage Axioms*". In: *The Quarterly Journal of Economics* 75.4, pp. 643–669. ISSN: 0033-5533. DOI: 10.2307/1884324. URL: https://doi.org/10.2307/1884324.

Evans, Owain, Andreas Stuhlmueller, and Noah D. Goodman (2015). *Learning the Preferences of Ignorant, Inconsistent Agents*. arXiv: 1512.05832 [cs.AI].

Frazier, David T., Robert Kohn, et al. (2023). *Reliable Bayesian Inference in Misspecified Models*. arXiv: 2302.06031 [stat.ME]. URL: https://arxiv.org/abs/2302.06031.

Frazier, David T., Christian P. Robert, and Judith Rousseau (Jan. 2020). "Model Misspecification in Approximate Bayesian Computation: Consequences and Diagnostics". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 82.2, pp. 421–444. ISSN: 1369-7412. DOI: 10.1111/rssb.12356. eprint: https://academic.oup.com/jrsssb/article-pdf/82/2/421/49321169/rssb12356-sup-0001-supinfo.pdf. URL: https://doi.org/10.1111/rssb.12356.

Frederick, Shane, George Loewenstein, and Ted O'Donoghue (2002). "Time Discounting and Time Preference: A Critical Review". In: *Journal of Economic Literature* 40.2, pp. 351–401. ISSN: 00220515. URL: http://www.jstor.org/stable/2698382 (visited on 08/04/2023).

Freedman, Rachel, Rohin Shah, and Anca Dragan (2020). "Choice Set Misspecification in Reward Inference". In: *IJCAI-PRICAI-20 Workshop on Artificial Intelligence Safety*. DOI: 10.48550/ARXIV.2101.07691. URL: https://arxiv.org/abs/2101.07691.

Givan, Robert, Thomas Dean, and Matthew Greig (2003). "Equivalence notions and model minimization in Markov decision processes". In: *Artificial Intelligence* 147.1. Planning with Uncertainty and Incomplete Information, pp. 163–223. ISSN: 0004-3702. DOI: https://doi.org/10.1016/S0004-3702(02)00376-4. URL: https://www.sciencedirect.com/science/article/pii/S0004370202003764.

Gleave, Adam et al. (2021). "Quantifying Differences in Reward Functions". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=LwEQnp6CYev.

Green, Leonard and Joel Myerson (1996). "Exponential versus hyperbolic discounting of delayed outcomes: Risk and waiting time". In: *American Zoologist* 36.4, pp. 496–505.

Griffin, Charlie et al. (2022). "All's Well That Ends Well: Avoiding Side Effects with Distance-Impact Penalties". In: *NeurIPS ML Safety Workshop*. URL: https://openreview.net/forum?id=3tgegVVh2j6.

Grünwald, Peter and Thijs van Ommen (2017). "Inconsistency of Bayesian Inference for Misspecified Linear Models, and a Proposal for Repairing It". In: *Bayesian Analysis* 12.4, pp. 1069–1103. DOI: 10.1214/17-BA1085. URL: https://doi.org/10.1214/17-BA1085.

Haarnoja, Tuomas et al. (Aug. 2017). "Reinforcement Learning with Deep Energy-Based Policies". In: *Proceedings of the 34th International Conference on Machine Learning*. Vol. 70. Proceedings of Machine Learning Research. Sydney, Australia: PMLR, pp. 1352–1361.

Hadfield-Menell, Dylan et al. (2016). "Cooperative Inverse Reinforcement Learning". In: *Advances in Neural Information Processing Systems*. Ed. by D. Lee et al. Vol. 29. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2016/file/c3395dd46c34fa7fd8d729d8cf88b7a8-Paper.pdf.

Haile, Philip and Elie Tamer (2003). "Inference with an Incomplete Model of English Auctions". In: *Journal of Political Economy* 111.1, pp. 1–51. ISSN: 00223808, 1537534X. URL: http://www.jstor.org/stable/10.1086/344801 (visited on 04/14/2025).

Hong, Joey, Kush Bhatia, and Anca Dragan (2022). *On the Sensitivity of Reward Inference to Misspecified Human Models*. arXiv: 2212.04717 [cs.LG].

Hussein, Ahmed et al. (Apr. 2017). "Imitation Learning: A Survey of Learning Methods". In: *ACM Comput. Surv.* 50.2. ISSN: 0360-0300. DOI: 10.1145/3054912. URL: https://doi.org/10.1145/3054912.

Ibarz, Borja et al. (Dec. 2018a). "Reward Learning from Human Preferences and Demonstrations in Atari". In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. NIPS'18. Red Hook, NY, USA: Curran Associates Inc., pp. 8022–8034. (Visited on 09/28/2023).

— (2018b). "Reward Learning from Human Preferences and Demonstrations in Atari". In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. Vol. 31. Montréal, Canada: Curran Associates, Inc., Red Hook, NY, USA, pp. 8022–8034.

Imbens, Guido and Charles Manski (May 2003). *Confidence intervals for partially identified parameters*. CeMMAP working papers 09/03. Institute for Fiscal Studies. DOI: 10.1920/wp.cem.2003.0903. URL: https://ideas.repec.org/p/azt/cemmap/09-03.html.

Jenner, Erik and Adam Gleave (2022). *Preprocessing Reward Functions for Interpretability*. arXiv: 2203.13553 [cs.LG].

Jeon, Hong Jun, Smitha Milli, and Anca Dragan (2020). "Reward-rational (implicit) choice: A unifying formalism for reward learning". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., pp. 4415–4426. URL: https://proceedings.neurips.cc/paper/2020/file/2f10c1578a0706e06b6d7db6f0b4a6af-Paper.pdf.

Jinnai, Yuu, David Abel, et al. (2019). *Finding Options that Minimize Planning Time*. arXiv: 1810.07311 [cs.AI]. URL: https://arxiv.org/abs/1810.07311.

Jinnai, Yuu, Jee Won Park, et al. (2019). *Discovering Options for Exploration by Minimizing Cover Time*. arXiv: 1903.00606 [cs.AI]. URL: https://arxiv.org/abs/1903.00606.

Jong, Nicholas and Peter Stone (Jan. 2005). "State Abstraction Discovery from Irrelevant State Variables." In: pp. 752–757.

Kahneman, Daniel and Amos Tversky (1979). "Prospect Theory: An Analysis of Decision under Risk". In: *Econometrica* 47.2, pp. 263–291. ISSN: 00129682, 14680262. URL: http://www.jstor.org/stable/1914185 (visited on 08/11/2022).

Karwowski, Jacek et al. (2023). *Goodhart's Law in Reinforcement Learning*. arXiv: 2310.09144 [cs.LG].

Kim, Kuno et al. (July 2021). "Reward Identification in Inverse Reinforcement Learning". In: *Proceedings of the 38th International Conference on Machine Learning*. Vol. 139. Proceedings of Machine Learning Research. Virtual: PMLR, pp. 5496–5505.

Kirby, Kris (Mar. 1997). "Bidding on the Future: Evidence Against Normative Discounting of Delayed Rewards". In: *Journal of Experimental Psychology: General* 126, pp. 54–70. DOI: 10.1037/0096-3445.126.1.54.

Kleijn, B.J.K. and A.W. van der Vaart (2012). "The Bernstein-Von-Mises theorem under misspecification". In: *Electronic Journal of Statistics* 6.none, pp. 354–381. DOI: 10.1214/12-EJS675. URL: https://doi.org/10.1214/12-EJS675.

Klein, Timo et al. (2023). *Active Third-Person Imitation Learning*. arXiv: 2312.16365 [cs.LG]. URL: https://arxiv.org/abs/2312.16365.

Knox, W. Bradley et al. (Mar. 2023). "Reward (Mis)Design for Autonomous Driving". In: *Artificial Intelligence* 316, p. 103829. ISSN: 0004-3702. DOI: 10.1016/j.artint.2022.103829. (Visited on 09/29/2023).

Krakovna, Victoria, Laurent Orseau, Ramana Kumar, et al. (2018). *Penalizing side effects using stepwise relative reachability*. DOI: 10.48550/ARXIV.1806.01186. URL: https://arxiv.org/abs/1806.01186.

Krakovna, Victoria, Laurent Orseau, Richard Ngo, et al. (2020). *Avoiding Side Effects By Considering Future Tasks*. DOI: 10.48550/ARXIV.2010.07877. URL: https://arxiv.org/abs/2010.07877.

Krakovna, Victoria, Jonathan Uesato, et al. (2020). *Specification gaming: the flip side of AI ingenuity*. https://deepmind.google/discover/blog/specification-gaming-the-flip-side-of-ai-ingenuity/.

Li, Lihong, Thomas Walsh, and Michael Littman (Jan. 2006). "Towards a Unified Theory of State Abstraction for MDPs." In.

Liao, Yuan and Anna Simoni (2019). "Bayesian inference for partially identified smooth convex models". In: *Journal of Econometrics* 211.2, pp. 338–360. ISSN: 0304-4076. DOI: https://doi.org/10.1016/j.jeconom.2019.03.001. URL: https://www.sciencedirect.com/science/article/pii/S0304407619300399.

Manheim, David and Scott Garrabrant (Feb. 2019). *Categorizing Variants of Goodhart's Law*. DOI: 10.48550/arXiv.1803.04585. arXiv: 1803.04585 [cs, q-fin, stat]. (Visited on 09/29/2023).

Manski, Charles (2003). *Partial Identification of Probability Distributions: Springer Series in Statistics*. English. Springer. ISBN: 9780387004549.

Mazur, JE (1987). "An adjusting procedure for studying delayed reinforcement (Vol. 5)". In: *Quant Anal Behav*, pp. 55–73.

Mccallum, Andrew Kachites and Dana Ballard (1996). "Reinforcement learning with selective perception and hidden state". AAI9618237. PhD thesis.

Metelli, Alberto Maria, Filippo Lazzati, and Marcello Restelli (2023). *Towards Theoretical Understanding of Inverse Reinforcement Learning*.

Michaud, Eric J., Adam Gleave, and Stuart Russell (2020). *Understanding Learned Reward Functions*. arXiv: 2012.05862 [cs.LG].

Moon, Hyungsik Roger and Frank Schorfheide (Apr. 2009). *Bayesian and Frequentist Inference in Partially Identified Models*. Working Paper 14882. National Bureau of Economic Research. DOI: 10.3386/w14882. URL: http://www.nber.org/papers/w14882.

Müller, Ulrich (Sept. 2013). "Risk of Bayesian Inference in Misspecified Models, and the Sandwich Covariance Matrix". In: *Econometrica* 81. DOI: 10.3982/ECTA9097.

Ng, Andrew Y, Daishi Harada, and Stuart Russell (1999). "Policy Invariance Under Reward Transformations: Theory and Application to Reward Shaping". In:

Proceedings of the Sixteenth International Conference on Machine Learning. Bled, Slovenia: Morgan Kaufmann Publishers Inc, pp. 278–287.

Ng, Andrew Y and Stuart Russell (2000). "Algorithms for Inverse Reinforcement Learning". In: *Proceedings of the Seventeenth International Conference on Machine Learning*. Vol. 1. Stanford, California, USA: Morgan Kaufmann Publishers Inc, pp. 663–670.

Orsini, Manu et al. (2021). "What Matters for Adversarial Imitation Learning?" In: *arXiv preprint* arXiv:2106.00672 [cs.LG]. To appear in *Proceedings of the 35th International Conference on Neural Information Processing Systems*, 2021.

Pan, Alexander, Kush Bhatia, and Jacob Steinhardt (2022). "The Effects of Reward Misspecification: Mapping and Mitigating Misaligned Models". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=JYtwGwIL7ye.

Pang, Richard Yuanzhe et al. (2022). "Reward Gaming in Conditional Text Generation". In: arXiv. DOI: 10.48550/ARXIV.2211.08714. (Visited on 09/29/2023).

— (2023). *Reward Gaming in Conditional Text Generation*. arXiv: 2211.08714 [cs.CL].

Paulus, Romain, Caiming Xiong, and Richard Socher (Feb. 2018). "A Deep Reinforced Model for Abstractive Summarization". In: *International Conference on Learning Representations*. (Visited on 09/29/2023).

Ramachandran, Deepak and Eyal Amir (2007). "Bayesian Inverse Reinforcement Learning". In: *Proceedings of the 20th International Joint Conference on Artifical Intelligence*. Hyderabad, India: Morgan Kaufmann Publishers Inc, pp. 2586–2591.

Ravindran, Balaraman and Andrew Barto (May 2003). "SMDP Homomorphisms: An Algebraic Approach to Abstraction in Semi-Markov Decision Processes". In.

Rothkopf, Constantin A and Christos Dimitrakakis (2011). "Preference Elicitation and Inverse Reinforcement Learning". In: *Machine Learning and Knowledge Discovery in Databases: ECML PKDD 2011, Proceedings, Part III*. Vol. 6913. Lecture Notes in Computer Science. Springer. Athens, Greece, pp. 34–48. DOI: 10.1007/978-3-642-23808-6-3.

Schlaginhaufen, Andreas and Maryam Kamgarpour (2023). *Identifiability and Generalizability in Constrained Inverse Reinforcement Learning*. arXiv: 2306.00629 [cs.LG].

Shah, Rohin, Noah Gundotra, Pieter Abbeel, and Anca Dragan (June 2019). "On the Feasibility of Learning, Rather than Assuming, Human Biases for Reward Inference". In: *Proceedings of the 36th International Conference on Machine Learning*. Vol. 97. Proceedings of Machine Learning Research. Long Beach, California, USA: PMLR, pp. 5670–5679.

Shah, Rohin, Noah Gundotra, Pieter Abbeel, and Anca D. Dragan (2019). *On the Feasibility of Learning, Rather than Assuming, Human Biases for Reward Inference*. arXiv: 1906.09624 [cs.LG].

Sharma, Pratyusha, Deepak Pathak, and Abhinav Gupta (2019). *Third-Person Visual Imitation Learning via Decoupled Hierarchical Controller*. arXiv: 1911.09676 [cs.LG]. URL: https://arxiv.org/abs/1911.09676.

Singh, Avi et al. (June 2019). "End-to-End Robotic Reinforcement Learning Without Reward Engineering". In: *Proceedings of Robotics: Science and Systems*. Freiburg im Breisgau, Germany. DOI: 10.15607/RSS.2019.XV.073.

Singh, Sumeet et al. (2018). *Risk-sensitive Inverse Reinforcement Learning via Semi- and Non-Parametric Methods*. arXiv: 1711.10055 [cs.AI].

Skalse, Joar and Alessandro Abate (June 2023a). "Misspecification in Inverse Reinforcement Learning". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 37.12, pp. 15136–15143. DOI: 10.1609/aaai.v37i12.26766. URL: https://ojs.aaai.org/index.php/AAAI/article/view/26766.

— (July 2023b). "On the limitations of Markovian rewards to express multi-objective, risk-sensitive, and modal tasks". In: *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*. Ed. by Robin J. Evans and Ilya Shpitser. Vol. 216. Proceedings of Machine Learning Research. PMLR, pp. 1974–1984. URL: https://proceedings.mlr.press/v216/skalse23a.html.

— (2024). *Quantifying the Sensitivity of Inverse Reinforcement Learning to Misspecification*. arXiv: 2403.06854 [cs.LG]. URL: https://arxiv.org/abs/2403.06854.

Skalse, Joar, Lucy Farnik, et al. (2023). *STARC: A General Framework For Quantifying Differences Between Reward Functions*. arXiv: 2309.15257 [cs.LG].

Skalse, Joar, Matthew Farrugia-Roberts, et al. (2022). "Invariance in Policy Optimisation and Partial Identifiability in Reward Learning". In: *arXiv preprint arXiv:2203.07475*.

Skalse, Joar, Niki Howe, et al. (2022). "Defining and Characterizing Reward Hacking". In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems*.

Stadie, Bradly C., Pieter Abbeel, and Ilya Sutskever (2019). *Third-Person Imitation Learning*. arXiv: 1703.01703 [cs.LG]. URL: https://arxiv.org/abs/1703.01703.

Sutton, Richard S and Andrew G Barto (2018). *Reinforcement Learning: An Introduction*. second. MIT Press. ISBN: 9780262352703.

Sutton, Richard S., Doina Precup, and Satinder Singh (1999). "Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning". In: *Artificial Intelligence* 112.1, pp. 181–211. ISSN: 0004-3702. DOI: https://doi.org/10.1016/S0004-3702(99)00052-1. URL: https://www.sciencedirect.com/science/article/pii/S0004370299000521.

Swamy, Gokul et al. (2021). *Of Moments and Matching: A Game-Theoretic Framework for Closing the Imitation Gap*. arXiv: 2103.03236 [cs.LG]. URL: https://arxiv.org/abs/2103.03236.

Thaler, Richard (1981). "Some empirical evidence on dynamic inconsistency". In: *Economics Letters* 8.3, pp. 201–207. ISSN: 0165-1765. DOI: https://doi.org/10.1016/0165-1765(81)90067-7. URL: https://www.sciencedirect.com/science/article/pii/0165176581900677.

Turner, Alex, Neale Ratzlaff, and Prasad Tadepalli (2020). "Avoiding Side Effects in Complex Environments". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., pp. 21406–21415. URL: https://proceedings.neurips.cc/paper/2020/file/f50a6c02a3fc5a3a5d4d9391f05f3efc-Paper.pdf.

Viano, Luca et al. (2021). "Robust Inverse Reinforcement Learning under Transition Dynamics Mismatch". In: *Advances in Neural Information Processing Systems*. Ed. by A. Beygelzimer et al. URL: https://openreview.net/forum?id=t8HduwpoQQv.

Vuorio, Risto, Mattie Fellows, et al. (2024). *A Bayesian Solution To The Imitation Gap*. arXiv: 2407.00495 [cs.LG]. URL: https://arxiv.org/abs/2407.00495.

Vuorio, Risto, Pim de Haan, et al. (2024). *Deconfounding Imitation Learning with Variational Inference*. arXiv: 2211.02667 [cs.LG]. URL: https://arxiv.org/abs/2211.02667.

Watanabe, Sumio (2009). *Algebraic Geometry and Statistical Learning Theory*. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press.

— (2018). *Mathematical Theory of Bayesian Statistics*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. CRC Press, Taylor & Francis Group. ISBN: 9781482238068.

Weihs, Luca et al. (2021). *Bridging the Imitation Gap by Adaptive Insubordination*. arXiv: 2007.12173 [cs.LG]. URL: https://arxiv.org/abs/2007.12173.

White, H (1982). "Maximum Likelihood Estimation of Misspecified Models." In: *Econometrica* 50.(1), pp. 1–25.

White, Halbert (1994). *Estimation, Inference and Specification Analysis*. Econometric Society Monographs. Cambridge University Press. DOI: 10.1017/CCOL0521252806.

Wulfe, Blake et al. (2022). *Dynamics-Aware Comparison of Learned Reward Functions*. arXiv: 2201.10081 [cs.LG].

Yamaguchi, Shoichiro et al. (May 2018). "Identification of animal behavioral strategies by inverse reinforcement learning". In: *PLOS Computational Biology* 14.5, pp. 1–20. DOI: 10.1371/journal.pcbi.1006122. URL: https://doi.org/10.1371/journal.pcbi.1006122.

Yang, Ziheng and Tianqi Zhu (2018). "Bayesian selection of misspecified models is overconfident and may cause spurious posterior probabilities for phylogenetic trees". In: *Proceedings of the National Academy of Sciences* 115.8, pp. 1854–1859. DOI: 10.1073/pnas.1712673115. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.1712673115. URL: https://www.pnas.org/doi/abs/10.1073/pnas.1712673115.

Zhuang, Simon and Dylan Hadfield-Menell (2020). "Consequences of Misaligned AI". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., pp. 15763–15773. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/b607ba543ad05417b8507ee86c54fcb7-Paper.pdf.

Ziebart, Brian D (2010). "Modeling Purposeful Adaptive Behavior with the Principle of Maximum Causal Entropy". PhD thesis. Carnegie Mellon University.